# Articulatory Phonetic Features for Improved Speech Recognition


Huang Guangpu

School of Electrical & Electronic Engineering

Nanyang Technological University

A thesis submitted for the degree of

*Doctor of Philosophy*

2012

Dedicated to friends and family.

# Acknowledgements

I sincerely thank Prof. Er Meng Joo, whose profound interest in automatic speech processing caused me to undertake this project from the school of EEE in NTU. A big thanks to Zhou Yong, who has offered so many advices on programming. I am also thankful for the authors whose articles and books have kept me company for these years. Their names are in the references. Finally I would like to my family and friends who have shown unconditional support during my graduate studies.

# Abstract

This thesis elaborates the use of speech production knowledge in the form of articulatory phonetic features to improve the robustness of speech recognition in practical situations. The main concept is that natural speech has three attributes in the human speech processing system, i.e., the motor activation, the articulatory trajectory, and the auditory perception. Consequently, the research work has three components. First, it describes an adaptive neural control model, which reproduces the articulatory trajectories and retrieves the motor activation patterns in a bio-mechanical speech synthesizer. Second, by manipulating the elastic vocal tract walls, the synthesizer produces the overall articulatory-to-acoustic trajectory map for English pronunciations. Third, the articulatory phonetic features are extracted in neural networks for speech recognition in cross-speaker and noisy conditions. The experimental results are compared with the traditional hidden Markov baseline system.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

**Roman Symbols**

| | |
|---|---|
| 2-D | 2-Dimensional |
| 3-D | 3-Dimensional |
| AF | Articulatory Feature |
| ANN | Artificial Neural Networks |
| API | Articulatory Phonetic Inversion |
| ASR | Automatic Speech Recognition |
| BBN | Bolt, Beranek and Newman Technologies |
| BFCC | Bark Frequency Cepstral Coefficient |
| CMU | Carnegie Mellon University |
| CNET | Centre national d'*é*tudes des t*élé*communications (National Center for Telecommunication Studies) |
| CV | Consonant Vowel |
| DARPA | Defense Advanced Research Projects Agency |
| DBN | Dynamic Bayesian Network |
| E-FNN | Extended Fuzzy Neural Network |
| EM | Expectation Maximization |

| | |
|---|---|
| EMA | Electromagnetic Articulograph |
| EMG | Electromyography |
| EPH | Equilibrium Point Hypothesis |
| FFT | Fast Fourier transform |
| FNN | Fuzzy Neural Network |
| GD-FNN | Generalized Dynamic Fuzzy Neural Network |
| GGa | Anterior Genioglossus |
| GGp | Posterior Genioglossus |
| HG | Hyoglossus |
| HMM | Hidden Markov Model |
| HTK | Hidden Markov Model Toolkit |
| IBM | International Business Machines Corporation |
| IPA | International Phonetic Alphabet |
| LI | Lower Incisor |
| LL | Lower Lip |
| MA | Masseter |
| MFCC | Mel Frequency Cepstral Coefficient |
| MIT | Massachusetts Institute of Technology |
| MLLR | Maximum Likelihood Linear Regression |
| MLP | Multi Layer Perceptron |
| MOA | Manner of Articulation |
| MOCHA | Multichannel Articulatory Corpus |

| | |
|---|---|
| MSD | Mass Spring Damper |
| MV | Motor Variable |
| NLU | Natural Language Understanding |
| OO | Orbicularisoris |
| PAF | Phonological Articulatory Feature |
| PCA | Principle Component Analysis |
| PDE | Partial Differential Equation |
| PID | Proportional Integral Derivative |
| PLP | Perceptual Linear Prediction |
| POA | Place of Articulation |
| RAT | Regional Articulatory Target |
| RBF | Radial Basis Function |
| RMSE | Root Mean Square Error |
| RNN | Recurrent Neural Network |
| RO | Risorius |
| ROS | Rate of Speech |
| SCRIBE | Spoken Corpus Recordings In British English |
| SG | Styloglossus |
| SL | Superior Longitudinal |
| SNR | Signal to Noise Ratio |
| TB | Tongue Body |
| TDNN | Time Delay Neural Network |

| | |
|---|---|
| TIMIT | Texas Instrument and Massachusetts Institute of Technology |
| TLM | Transmission Line Model |
| TR | Tongue Root |
| TT | Tongue Tip |
| TTS | Text to Speech |
| TV | Tract Variable |
| UL | Upper Lip |
| VOT | Voice Onset Time |
| VTL | Vocal Tract Length |
| VTLN | Vocal Tract Length Normalization |
| WER | Word Error Rate |
| WSJ | Wall Street Journal |

# Chapter 1

# Introduction

## 1.1 Motivation

Human-machine communication plays a dominant rule in productivity in the information age. Speech is the most convenient way of communication for human. One distinct advantage of speech is its capacity and efficiency. Spontaneous speech conveys 2.0 to 3.6 words per second on average. In contrast, average computer user can only type about 0.2 to 0.4 words per second, while the most skilled typists only achieve 1.6 to 2.5 words per second. In the actual working environment, that is thinking whilst typing, speech gives 3 words per second, while typing only 0.3 words, ten times slower (Martin and Jurasfsky, 2008). The use of speech will thus free the hands from the keyboard for multi-tasking. It will bring as many advantages as when the hands were first freed from the ground when human learned to walk.

Digital speech processing consists of two main components: automatic speech recognition (ASR), and natural language understanding. ASR transcribes text from the spoken input. Language understanding extracts the meaning of the written or spoken text and gives feedback accordingly (Woodland, 1998). A simple information retrieval system involving the processes is shown in Fig. 1.1, which distinguishes the two processes in the dashed box. Speech research encompasses a broad range of technical challenges, including automatic recognition of words and phrases in the speech signal, extraction of keywords or key-phrases in the

recognized utterances, and understanding of the spoken utterances. It has already shown great promises in many areas that are beneficial for the public, for example, dictation software in personal computers, automatic call routing services, telephone inquiry systems for share prices, train timetables and banking services. AT& T's VRCP, a five-word keyword spotting system, which automates billions of calls every year saves operating cost in the scale of hundreds of million of dollars (Cox, 2000). Other applications in darkroom operations have also greatly improved the productivity of industries. It is perceivable that the advent of powerful computing devices and the fast improvements of microprocessors will continue to bring more important scientific advances in speech technologies that benefit human society (Juang and Furui, 2000; Woodland, 1998).

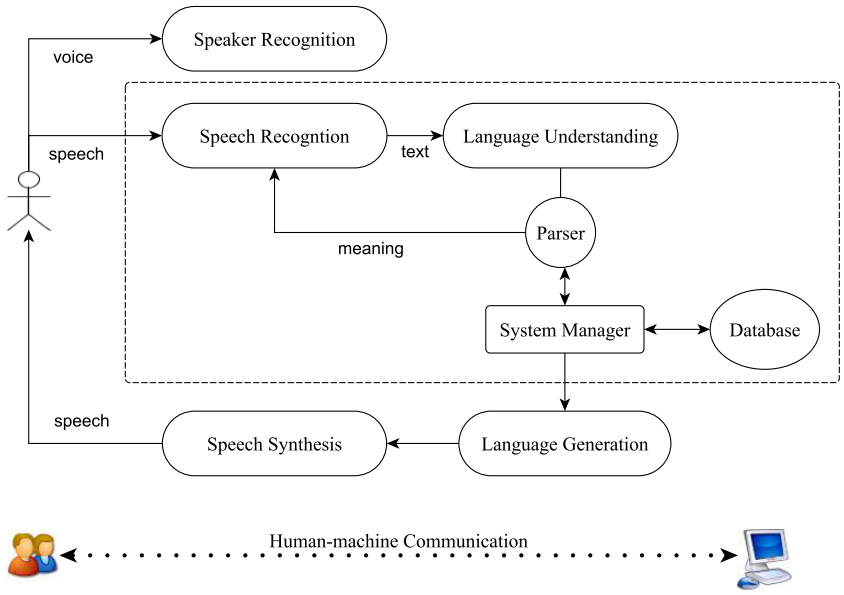



Figure 1.1: Human-computer interaction of the information retrieval system.

## 1.2 Objectives

This research aims to use the articulatory phonetic features to improve the accuracy and robustness of speech recognition in practical situations. It focuses on a pronunciation modeling method that combines the three attributes of English

speech: the motor activation, the articulatory trajectory, and the auditory perceptions. The method is analogous to that of the human speech processing system. The hypothesis is that the phonemes exhibit distinctive characteristics in the three different feature spaces. The author's research work has three main goals. First, it aims to design an adaptive neural control model for a bio-mechanical speech synthesizer. The synthesizer should reproduce the articulatory trajectories and retrieve the motor activation patterns in the speech sound. Second, it aims to use the bio-mechanical synthesizer to map the articulatory-acoustic trajectories. Third, it aims to derive the set of articulatory phonetic features for English speech. These features serve to embed additional knowledge sources in speech recognition systems.

### 1.2.1 Robustness in ASR

#### 1.2.1.1 Speech Variability Sources

ASR remains a challenging problem due to the variability of the speech signals. Speech conveys a message with multiple levels of knowledge sources, e.g., discourse, semantics, syntax, phonological, phonetic, acoustic, and articulatory. It also conveys information about the speaker such as gender, age, social status, geographical origin, health status, emotional state, and voice identity. The sources of variability can be generally classified as the following:

1. Intra-speaker (same speaker) variability,
   (a) Speaker physiology
   (b) Language proficiency
2. Inter-speaker (cross speaker) variability,
3. Linguistic variability,
   (a) Speaking style
   (b) Disfluency
   (c) Rate of speech

   (d) Co-articulation

4. Channel variability.

   (a) Background noise

   (b) Channel noise

   (c) Room reverberation

The ideal ASR systems must cope with these sources of variability to achieve high accuracy and robustness. Current systems are usually compromised the goal by specifying their constrains (Benzeghiba et al., 2007; Woodland, 1998),

1. vocabulary size,

2. mode of speech: isolated words versus continuous speech,

3. speakers dependence,

4. style of speech: read versus spontaneous.

These include the isolated word recognition (system 1 in Table 1.1), connected word recognition (system 2), conversational speech recognition (system 3), and conversational speech understanding (system 4) systems [1]. A number of benchmarking databases have also been constructed. For example, the DARPA resource

[1]

**MIT** Massachusetts Institute of Technology

**IBM** International Business Machines Corporation

**CMU** Carnegie Mellon University

**HTK** Hidden Markov Model Toolkit

**BBN** Bolt, Beranek and Newman Technologies

**CNET** Centre national d'études des télécommunications (National Center for Telecommunication Studies)

management (RM) database, Texas Instrument (TI)46 - isolated digits, TIDIGITS - connected digits, Alpha-Numeric (AN)4 - 100 words vocabulary, Texas Instrument and Massachusetts Institute of Technology (TIMIT), Wall Street Journal (WSJ)5K - 5,000 words vocabulary, and WSJ20K - 20,000 words vocabulary. Table 1.1 lists the groups of speech recognition and understanding systems according to the constraints.

Table 1.1: Types and performance of speech recognition and understanding systems. Detailed analysis and comparisons of these systems can be found in (Klatt, 1977; Reddy, 1976; Siroux and Gillet, 1985).

| | Mode of Speech | Vocabulary Size | Software Systems | Word Accuracy |
|---|---|---|---|---|
| 1. | isolated words | $10-30$ | dictation systems | $98-99.8\%$ |
| 2. | connected words | $10-500$ | Bell Lab: telephone voice-operation | $\leq 96\%$ |
| | | | IBM: voice-activated typewriter | $\leq 95\%$ |
| 3. | read speech | $30-2000$ | CMU: Hearsay-I, Sphinx & Dragon | $\leq 83\%$ |
| | | | MIT: Lincoln system | $\leq 90\%$ |
| | | | IBM: ViaVoice | $\leq 97\%$ |
| 4. | conversational speech | $100-5000$ | CMU: Hearsay-II, Harpy | $\leq 65\%$ |
| | | | BBN: Hwim, Speechlis | $\leq 75\%$ |
| | | | CNET: KEAL | $\leq 65\%$ |

#### 1.2.1.2 ASR Diagnosis

ASR technology has improved remarkable in the past six decades. However, compared to human performance, it still has a long way to go. The major concern is that the human listener outperforms the most advanced speech recognition system not only in noisy conditions but also in quiet environment. Even when the training provides acoustic/language models which are almost perfectly matched to the testing conditions, the ASR system still fail match to its human counterpart (Sroka and Braida, 2005). There is a deficiency gap of 10% versus 1% word error rate (WER) for the WSJ task between the human and the machine performance.

There are two levels of differences between the ASR system and the human listener. First the single microphone remains a common input device in most ASR applications. Yet human has two ears that allow directional hearing, localization, and tracking. Though the microphone arrays have been used to resemble

this effect (Che et al., 1994; Shimizu et al., 2000), the human ear is much more sophisticated than the microphone both anatomically and functionally. Current recording devices and feature extraction methods in ASR are not as competent as the human ears (Parham et al., 2006). Second the human brain consists of approximately 100 billion neurons with a vast around of interconnection, whereas the most powerful computer processors at present have less than 100 million transistors, about one thousand fold less in the basic computing units (Martin and Jurasfsky, 2008). Moreover, a transistor is not the same as a neuron. A neuron is a non-linear summation of inputs working in an inherently analog fashion, whereas a transistor, only mimics such analogous behaviors in a digital form, with only two outputs, 0 and 1 (Parham et al., 2006). In the end, the robustness and accuracy of human speech recognition would probably be more than the functionality of either the brain or the ears. The human speech recognition may in fact rely on cooperating both in the central nervous system.

### 1.2.2 Speech Production Knowledge in ASR

Conventional ASR system treats human speech as a concatenation of acoustic observations to allow probabilistic modeling of the phonetic sequences, e.g., using the hidden Markov models (HMMs) . The approach becomes problematic when dealing with the variabilities of natural conversational speech. In recent years, many articulatory and auditory based processing methods have been proposed to address the problem of phonetic variations in a number of frame-based, segment-based, and acoustic landmark systems (King et al., 2007; Stevens, 2002). For instance, the direct articulatory data have been collected through the use of electromagnetic articulograph (EMA) , X-ray analysis, and laryngograph (i.e., electroglottograph). They provide good references as well as additional knowledge sources for physiological speech studies (Richmond, 2009). The so-called articulatory features (AFs) have also improved the recognition performance of many ASR systems. For example, AFs derived from phonological rules have outperformed the acoustic HMM baseline in a series of phoneme recognition tasks (King et al., 2007; Kirchhoff et al., 2002; Saenko et al., 2005). Similarly, experimental studies of the mammalian peripheral and central auditory organs have

also introduced many perceptual processing methods. For example, several auditory models have been constructed to simulate human hearing, e.g., the ensemble interval histogram, the lateral inhibitory network, and Meddis' inner hair-cell model (Holmberg et al., 2006; Jankowski et al., 1995; Jeon and Juang, 2007). Auditory based features such as mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive coefficients are now widely used in ASR systems (Hermansky and Morgan, 1994; Holmberg et al., 2006).

In addition, the discovery of mirroring neurons suggests the bi-directionality of human speech production and perception. It urges researchers to investigate both aspects collectively to break the bottleneck of current speech recognition systems (Levelt, 1999). Previously, Guenther et al. (2006) have constructed a neural model to organize the accumulated pool of articulatory and acoustic data in a framework of humanoid sensory blocks. Following their footsteps, Kröger et al. (2009) built a neural computation model to enable parallel production and perception of simple syllables such as vowels, consonant vowel (CV), vowel consonant vowel (VCV), and consonant vowel consonant (CVC) patterns. However, up till now, the acoustic and the articulatory features have mostly been used as additional input streams or as internal representations in conventional ASR systems such as HMMs (Siniscalchi and Lee, 2009), multi-layer perceptrons (MLPs) (Kirchhoff et al., 2002), time-delay neural networks (TDNNs) (Schuster and Paliwal, 1997), radial basis function-based neural networks (RBF-NNs) (Yousefian et al., 2008), dynamic Bayesian networks (DBNs) (Frankel et al., 2007), and their hybrids (King et al., 2007; Trentin and Gori, 2001).

The main difficulty of using the multiple speech attributes in the ASR systems is the non-linearity between the articulatory data and the acoustic data. On the one hand, the articulatory-to-acoustic mapping is not one-to-one. In other words, there are more than one vocal shapes that can produce the same speech sound. On the other hand, both the articulatory and the acoustic data contain variations for the same speech sound, where the exact pronunciation depends on the context and the speaker. The acoustic-to-articulatory mapping, also known as the inverse mapping, remains a difficult problem in speech research (Toda et al., 2008).

## 1.3 Major Contribution of Thesis

ASR has many limitations in practical applications. The thesis summarizes the author's research work on using articulatory features for improved speech recognition. Major contribution of the thesis include:

1. Different from the existing phonological articulatory features (PAFs) which are derived from the broad linguistic definitions, e.g., manner of articulation (MOA) and place of articulation (POA), as used in (Frankel et al., 2007; Kirchhoff et al., 2002; Siniscalchi and Lee, 2009), a more reliable heuristic mapping strategy is used to retrieve a set of articulatory phonetic features (APFs) for the pronunciation models on a set of hand-labeled sentences. In addition, a heuristic learning algorithm is used to embed two knowledge based rules: the listener-oriented maximization of auditory discriminations from human speech perception and the speaker-oriented minimization of articulatory effort from human speech production.

2. This study uses a neural based articulatory phonetic inversion (API) model to find the abstract phonetic representation for improved speech recognition. The approach roots in the concept that the speech sound occupies a spread region, rather than isolated points, in the auditory and the articulatory domain (Damper and Harnad, 2000; Kielar et al., 2011; Mottonen and Watkins, 2009). What differs this study from others is the *unified explanation* of speech events in the production and the perception domains. The proposed pronunciation modeling method distinguishes the base-forms from the variations of English phonemes using multiple knowledge sources that are not present in conventional classifiers.

3. It addresses the non-uniqueness and the non-linearity issues in the inversion experiments by incorporating the multiple knowledge sources at three places.

    - In the control model, the bio-mechanical synthesizer approximates the human anatomy in physiological and functional properties.

- In the pronunciation model, the heuristic learning algorithm mimics the experience of human speech acquisition and production.
- In the inversion model, the data clustering algorithm minimizes the within-class scatter distance and maximizes the across-class scatter distance in the synthetic data, which is analogous to the categorical nature of human speech perception.

4. The research work focuses on finding and resolving the bottleneck of current speech processing techniques to improve the accuracy and the robustness. It reports the frame level accuracy and identifies the phoneme error patterns in ASR systems.

## 1.4 Organization of Thesis

The research method in this thesis is analogous to the human speech processing. It aims to simultaneously model the three attributes of human speech: the motor activation (speech signals in the brain), the articulatory trajectory (in the vocal tract), and the auditory perceptions (in the ear). The technical chapters, Chapter 3, 4, and 5, support the research goal, which aims to use the articulatory phonetic features to improve the accuracy and robustness of speech recognition. The hypothesis is that the phonemes exhibit distinctive characteristics in the different feature spaces. The rest of the thesis is organized as follows.

Chapter 2 reviews the ASR techniques in the literature.

Chapter 3 focuses on motor activation. It presents the adaptive neural control scheme based on fuzzy logic and neural networks. The proposed controller tracks the articulatory movements of the human vocal tract and infers the activation patterns of the underlying muscular structures. It is able to manipulate the mass-spring based elastic tract walls in a 2-D articulatory synthesizer to realize efficient speech motor control and to generate the articulatory-acoustic map of English phonemes.

Chapter 4 focuses on speech production and perception. It presents the mapping between the articulatory trajectories and the auditory parameters. By analyzing the multi-dimensional articulatory-acoustic attributes, this chapter shows

that the articulatory feature space presents a much smaller variance than the acoustic feature space. The phonetic analysis uses the non-uniform segments, i.e., the phonetic base-forms, the broad phoneme forms, and the narrow phonetic forms, to represent the variations of English pronunciations. Each segment can have more than one narrow phonetic labels to account for the source of variations such as the co-articulation effects in conversational speech.

Chapter 5 focuses on deploying the above method in machine-based speech recognition. First it retrieves the articulatory data from the acoustic data through a neural inversion module. Second it uses the set of auditory and articulatory features in a neural recognition module. The speech recognition experiments are carried out to test the accuracy and the robustness of the proposed technique dealing with the speaker variation and the noise contamination.

Chapter 6 concludes the thesis report, and recommends directions for future researches.

# Chapter 2

# Literature Review

## 2.1 Overview

This chapter reviews the techniques in automatic speech recognition (ASR). Section 2.2 presents the theoretical basis of human speech processing, including the production (2.2.1) and the perception (2.2.2). Section 2.3 introduces the digital representations and properties of the speech signal. Section 2.4 reviews the two types of speech recognizers, including the statistical method (2.4.1) and the connectionist method (2.4.2). Section 2.5 reviews the robustness techniques dealing with noise contamination (2.5.1) and speaker variations (2.5.2), and the use of articulatory cues (2.5.3).

Besides authorized books by linguists and speech experts, the primary sources of information in this review are IEEE Transactions of Acoustic, Speech, and Signal Processing, Journal of the Acoustical Society of America, Computer Speech and Language, Speech Communication, and Computational Linguistics. Major conferences in speech processing such as the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and Interspeech are also covered. Interested readers can refer to these publications for more information.

## 2.2 Human Speech Processing

Human language, either verbal speech or written text, is in many ways a creative activity, which distinguishes from animal communication mainly in two aspects. First the human listener is able to understand an indefinite number of new expressions across speakers and contextual variations after mastering a language. Second the human speaker is able to produce an equally indefinite number of expressions which can also be understood by others who share this ability (Chomsky, 2006). Accordingly, the nature of human speech encloses two folds. First speech exists as an information carrier in certain transmission channels, e.g., the sound waveform in the air (Allen, 1996). Second speech exists as part of the intellectual heritage which is dependent upon the influence of social communications and instructions through intuition, active learning, and experiences. Thus speech recognition can be outlined as a mirroring chain of events, as shown in Fig. 2.1. The left side shows the production fold where a message is encoded with articulatory gestures, and the right side shows the perception fold where the auditory signals are decoded to retrieve the message.

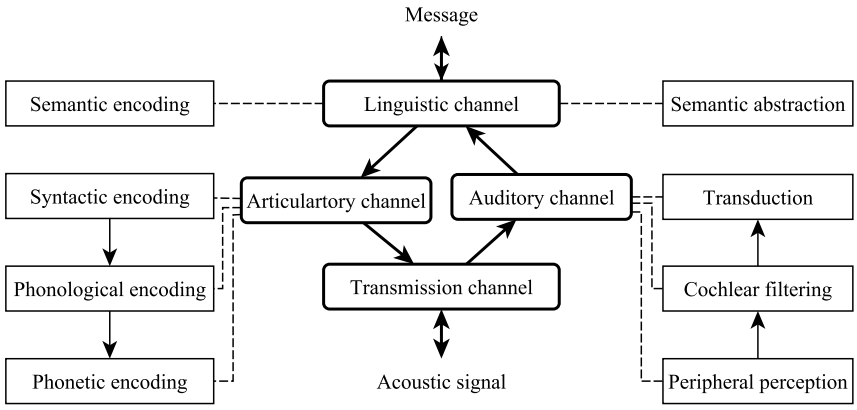



Figure 2.1: Overview of human speech production (left) and perception (right) in the mirroring chain of events.

### 2.2.1 Production Basis

There are a number of traditions and schools of thoughts regarding the theory of speech production, which generally consists of four stages (Martin and Jurasfsky, 2008), as illustrated on the left hand side of Fig. 2.1:

1. semantic encoding: it forms the concept and the intended message;
2. syntactic encoding: it specifies the syntactic frame or abstract form of words (without sound information);
3. phonological encoding: it retrieves the sound properties, i.e., phonological rules, for the phonemes and the syllables;
4. phonetic encoding: it articulates the sound through the glottis and the vocal tract using the appropriate motor coordination.

The semantic encoding stage is usually concerned with natural language understanding, though it is also sometimes used in ASR for specified tasks, e.g., small vocabulary digit and name dictations. This study will focus on the last three stages, where the interaction of different physiological structures in the articulatory space transforms the aerodynamic and myoelastic energy into acoustic signals. Speech is first created with pulmonary pressure provided by the lungs that generates sound by the glottal excitation in the larynx. Then it is modified by the vocal tract into different vowels and consonants. The vocal tract is an acoustic tube located between the vocal cords and the lips. A secondary tube, the nasal tract separates from the first by the velum. The shape of the vocal tract is determined by the position of the lips, jaw, tongue,and velum. Sound is generated by the different combination of the above apparatus. For example, voiced sounds are produced by the vibration of the vocal cords. Fricative sounds are produced by the constriction at different places in the vocal tract. Plosive sounds are produced by completely closing the vocal tract, building up the pressure before quickly releasing it (Schafer and Rabiner, 1975).

### 2.2.2 Perception Basis

Psycholinguistic studies deal with human speech perception at four stages (Jeon and Juang, 2007; Slaney, 1998), as shown on the right hand side of Fig. 2.1:

1. Peripheral perception: the outer ear receives the acoustic stimuli on a non-linear frequency scale, i.e., frequency selectivity;

2. Cochlear filtering: the cochlea maps the incoming tones to various spatial locations on the tonotopic axis of the central auditory system on the basilar membrane;

3. Transduction: the mechanical displacements along the basilar membrane converts into electrical activities on the topographically ordered array of auditory nerve fibers;

4. Semantic abstraction: the organ of corti transmits the electric pulses, which literally contain all the acoustic information, through the vestibulo-cochlea nerve to the primary auditory cortex on the brain atmospheres.

The goal of ASR is ultimately that of a court-recorder, which transcribes speech to text (Martin and Jurasfsky, 2008). This study will focus on the first three stages where the stimulated sound waves are transmitted to distinctive excitation patterns. For auditory feature extraction, the bark-frequency cepstral coefficients (BFCCs) are used to approximate the excitation patterns of the auditory nerve fibers. The ongoing debate is whether speech perception is necessarily a passive receptive task. Simply put, the lower level auditory stages may not be able to act independently to identify the phonemes without higher level sources such as morphology, syntax and semantics (Pisoni and Remez, 2004). However, human hearing experiments have repeatedly shown the determinant role of the lower level acoustic cues such as voicing onset time (VOT) in voicing detection and phoneme recognition (Zue, 2004). Categorical perception is also observed in infant studies, where babies with no prior linguistic knowledge tend to ignore within-class differences and enforce cross-class contrast during speech acquisition (Boets et al., 2007). Thus, it is feasible to assume that the auditory chain can generate phoneme strings, or approximation of such, e.g., phone classes, independent of higher level semantic knowledge.

## 2.3 Digital Speech Processing

The digital representation of speech includes the time-domain and the frequency-domain methods.

### 2.3.1 Time-domain Analysis

Human speech is composed of short quasi-stationary intervals of about 10 to 30 ms duration, when the characteristics of the waveform remain roughly invariant (Schafer and Rabiner, 1975). The concept of short-time analysis is the key to obtain adequate parametric representations of the speech. For a real discrete-time signal $x(n)$, its energy is,

$$E(n) = \sum_{n=-\infty}^{\infty} x^2(n) \tag{2.1}$$

For the non-stationary speech signals, it is more appropriate to calculate the energy as the following,

$$E(n) = \sum_{m=0}^{N-1} w(m)x(n-m)^2 \tag{2.2}$$

where $N$ samples of $x(n)$ is selected through a weighing window $w(m)$. It can be viewed as the sequence $x^2(n)$ being filtered by a finite impulse response filter $w^2(n)$. The choice of window function $w(m)$ and segment length $N$ determines the quality and accuracy of the signal energy measurement. $E(n)$ separates voiced speech segments from unvoiced ones.

Unvoiced sounds usually have much lower energy level than voiced sounds. Using autocorrelation, the internal structure of the signal can be displayed. The autocorrelation function of a discrete-time signal $x(n)$ is defined as,

$$\varphi(m) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+m). \tag{2.3}$$

For periodic voiced speech the autocorrelation shows the periodicity of the segments,

$$\varphi(m) = \varphi(m+P). \tag{2.4}$$

For non-periodic unvoiced speech and noise the function has a sharp peak at $m = 0$ and falls flat rapidly as $m$ increases. During speech processing the function is often modified to suit the quasi-stationary property of the signal. The revised short-time autocorrelation function is defined as

$$\varphi_l(m) = \frac{1}{N} \sum_{n=0}^{N'-1} x_l(n) x_l(n+m), \qquad 0 \leq m \leq M_0 - 1 \tag{2.5}$$

where $x_l(n) = x(n+l)$ for $0 \leq n \leq N-1$, $M_0$ is the maximum observation range inclusive of at least two periods of the signal. However, the direct computation of $\varphi_l(m)$ for $0 \leq m \leq M_0 - 1$ has a complexity proportional to $M_0 \cdot N$, which can be a significant overhead factor.

### 2.3.2 Frequency-domain Analysis

#### 2.3.2.1 Short-time Apectrum Analysis

For the quasi-stationary characteristics of speech, fast Fourier transform (FFT) produces the spectrum. The discrete short-time spectrum of $x(n)$ is defined as,

$$X_l(\omega) = \sum_{n=-\infty}^{l} x(n) h(l-n) e^{-j\omega n} \tag{2.6a}$$

$$= |X_l(\omega)| \, e^{j\theta_l(\omega)} \tag{2.6b}$$

$$= a_l(\omega) - j b_l(\omega). \tag{2.6c}$$

One interpretation of (2.6) is that $X_l(\omega)$ is the Fourier transform of a sequence $x(n)$ weighted by a windowing function $h(l-n)$. Another way to understand the formula is the filter method, i.e., $h(n)$ as the impulse response of a low-pass digital filter with input $x(n)e^{-j\omega n}$ and output $X_l(\omega)$ at certain frequency $\omega$ (2.6b), which is also a combination of two filters with response $a_n(\omega)$ and $b_n(\omega)$ as in (2.6c). The short-time spectrum is defined as

$$X_l(\omega) = \sum_{n=0}^{N-1} x_l(n) w(n) e^{-j\omega n} \tag{2.7}$$

where $x_l(n) = x(n+l), n = 0, 1, \cdots, N-1$, and $l = 0, L, 2L, \cdots$.

The frequency resolution of the spectrum is proportional to the window length $N$. Large window size gives better and more complete knowledge about pitch and vocal tract transfer function information, but at the price of heavier computational load. There are a variety of methods for estimating fundamental frequency as well as other parameters such as the resonance frequencies or formants from the spectrum (Kinjo and Funaki, 2006; Schafer and Rabiner, 1975).

#### 2.3.2.2 Cepstrum Analysis

The spectrum of the FFT spectrum is the cepstrum, a play of words (s-p-e-c to c-e-p-s). Speech spectrum is the convolution of the excitation and the vocal tract impulse response based on the principle of superposition. Cepstrum takes the logarithm of the magnitude of the Fourier transform, i.e., the spectrum of the windowed signal, of the excitation and the vocal tract impulse response, sums the logarithms, and applies the inverse discrete Fourier transform. These operations transform convolution into addition.

Since the cepstrum of a periodic train of impulses will also be a train of impulses with the same spacing as the input, the cepstrum serves as an excellent basis for estimating the fundamental period of voiced speech and for determining where a particular speech segment is voiced or unvoiced (Schafer and Rabiner, 1975). Moreover, the cepstrum is able to show separately the non-overlapped vocal tract and excitation components. The former is often called the spectrum envelope, which can be obtained easily by linear filtering, e.g. the fast convolution method (Rabiner and Schafer, 1978; Ramirez, 2004; Schafer and Rabiner, 1975). In the cepstrum the presence of a strong peak indicates voiced signal. The location of the strong peak presented indicates the pitch period. The smoothed envelope indicates the vocal tract resonances or formant frequencies (Ali et al., 2006; Pitton et al., 1996).

#### 2.3.2.3 Perceptual Measures

Besides Hertz (Hz), there are two other scales: the perceptual mel-scale and the psycho-acoustic Bark-scale. as used in the mel-frequency cepstral coefficients

(MFCCs) and the perceptual linear prediction (PLP) features. Both are convertible with the Hz measure. Convert $f$ Hz to Mel-scale,

$$m = 2595log_{10}(\frac{f}{700} + 1) = 1127ln(\frac{f}{700} + 1), \tag{2.8}$$

and the inverse,

$$f = 700(10^{m/2595} - 1) = 700(e^{m/1127} - 1). \tag{2.9}$$

Convert $f$ Hz to Bark-scale,

$$\text{Bark} = 13arctan(0.00076f) + 3.5arctan((\frac{f}{7500})^2), \tag{2.10}$$

or

$$result = \left[\frac{26.81f}{1960 + f}\right] - 0.53, \tag{2.11}$$

and

$$\text{Bark} = \text{result} + \begin{cases} 0.15 * (2 - \text{result}), & \text{if result} < 2 \\ 0, & \text{if } 2 \leq \text{result} \leq 20.1 \\ 0.22 * (\text{result} - 20.1), & \text{if result} > 20.1, \end{cases} \tag{2.12}$$

and the inverse,

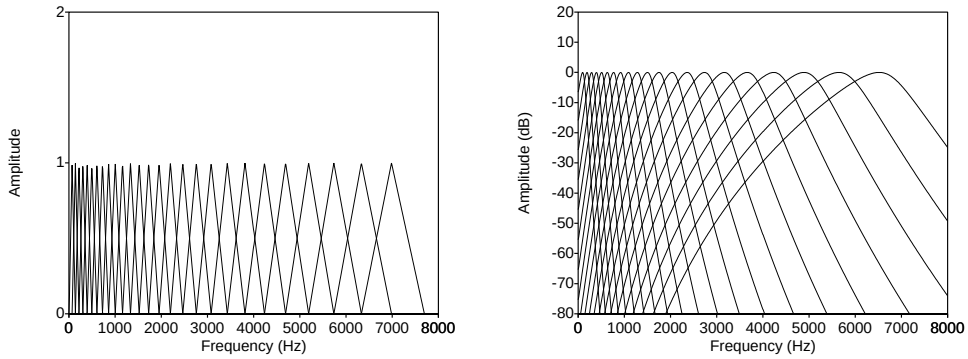
$$f = 52548/(z^2 - 52.56z + 690.39) \tag{2.13}$$

with $z$ in Bark-scale. Fig. 2.2 shows side by side the mel and the Bark scale filter banks. The following settings are used,

- Window length: 15 ms
- Time step: 5 ms
- Position of first filter: 1 Bark/100 mel
- Distance between filters 1 bark/100 mel

## 2.4 Automatic Speech Recognition

There are mainly two types of speech recognizers: the statistical method and the connectionist method.




(a) Mel filter banks (b) Bark filter banks

Figure 2.2: Illustration of the mel and the Bark scale filter banks.

## 2.4.1 Statistical Approaches

### 2.4.1.1 Hidden Markov Model

The most popular statistical method is the hidden Markov model (HMM), which has produced many powerful speech recognition engines such as Julius, the hidden Markov model toolkit (HTK), and Sphinx (Fry, 1959; Klatt, 1977; Lesser et al., 1975; Rabiner, 1989). It applies a finite-stated Markov model to estimate the output distributions. An HMM is defined by,

1. $N$, the number of individual states $S = S_1, S_2, \cdots, S_N$ with the state at time $t$ as $q_t$.

2. $M$, the number of distinct observation symbols per state $V = v_1, v_2, \cdots, v_M$.

3. The state transition probability distribution $A = a_{ij}$, where

$$a_{ij} = P\left(q_{t+1} = S_j | q_t = S_i\right), \qquad 1 \leq i, j \leq N. \tag{2.14}$$

4. The observation symbol probability distribution in state j, $B = b_j(k)$, where

$$b_j(k) = P\left(v_k \text{at } t | q_t = S_j\right), \qquad 1 \leq j \leq N; 1 \leq k \leq M. \tag{2.15}$$

5. The initial stat distribution $\pi = \pi_i$, where

$$\pi_i = P\left(q_1 = S_i\right) \qquad 1 \leq i \leq N. \tag{2.16}$$

Given appropriate values of $N, M, A, B$, and $\pi$, the HMM gives an observation sequence

$$O = O_1 O_2 \cdots O_T. \tag{2.17}$$

Generally an HMM model $\lambda$ is given in the compact form of

$$\lambda = (A, B, \pi) \tag{2.18}$$

to indicate the complete parameter set of the model. HMMs represent an effective and competitive learning paradigm. New data can be collected and learned during training to find the optimal estimates of $A$,$B$, and $\pi$, using, for example, the maximum-likelihood (ML) criterion. Three parameters need to be calculated (Rabiner, 1989),

1. the likelihood of a sequence of observations given a specific HMM, i.e., $P(O|\lambda)$,
2. the best sequence of model states, $Q = q_1, \cdots, q_6$,
3. the adjustment of model parameters $(A, B, \pi)$.

The first parameter scores how well a given model matches the given observation such as word sequences. It is usually solved by the forward-backward algorithm which sums over all possible state sequences to find the overall probability. The second parameter looks for the optimal state sequence $Q$ for the observation sequences. A formal technique addressing this problem, based on dynamic programming (DP) methods, is called the viterbi search which looks for the most likely state sequence. The third parameter trains the HMM for optimal global fitting, which is the most difficult part (Rabiner, 1989). Most HMM training algorithms use an iterative procedure to approximate global maximization, e.g., the expectation-maximization (EM) algorithm and the gradient techniques. The well known Baum-Welch algorithm is one example of the forward-backward algorithm, and a special case of the EM algorithm (Rabiner, 1989; Trentin and Gori, 2001; Woodland, 1998). Detailed information about HMM can be found in (Rabiner, 1989).

Fig. 2.3 shows a simple left-to-right Markov model with 6 states and 6 observations. HMM-based speech recognizers need to find the optimal $\lambda$ from the

training data based on the observation sequence $O = O_1 O_2 \cdots O_6$, e.g., acoustic features like MFCCs. The probability $P(O|\lambda)$ in the given model is obtained by summing the joint probability over all possible state sequences $q$,

$$P(O|\lambda) = \sum_{\text{state sequence}} P(O|Q,\lambda)P(Q|\lambda) \tag{2.19}$$

$$= \sum_{q_1,q_2,\cdots,q_6} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_5 q_6} b_{q_6}(O_6). \tag{2.20}$$

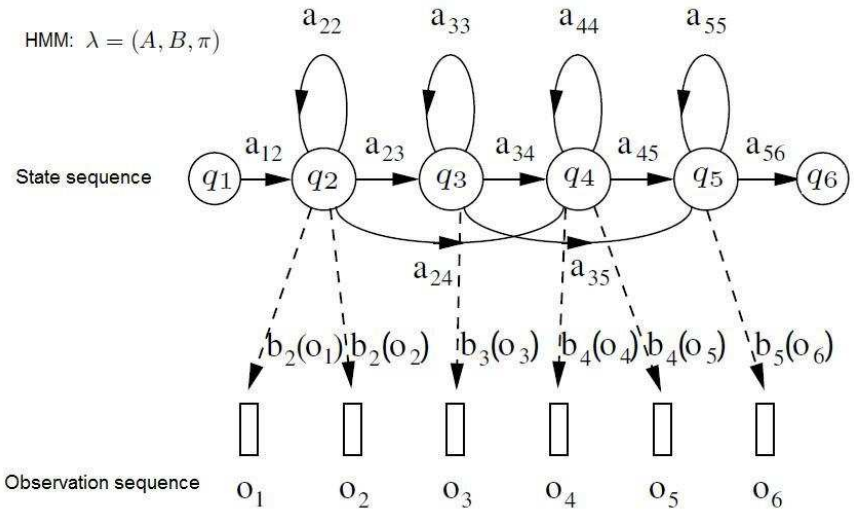



Figure 2.3: The Markov generation model: an example.

#### 2.4.1.2 HMM-based Speech Recognition

Fig. 2.4 illustrates the mathematical model of HMM-based speech recognition. The human speaker intends to deliver a message M. The central processor, i.e., the brain, first forms the message by a sequence of words W. W is further transformed into a sequence of phonetic sounds S. S radiates from the vocal tract and propagates in the air as Y. Y can be interpreted by the human listener, the ear. In ASR, Y is converted by a microphone into electric signals. Essentially HMM-based ASR system reverts the process to retrieve the essential message M. It has three components: acoustic models of the speech unit, usually phones, a word

models (also referred to as the lexicon or pronunciation dictionary) with pronunciation entries for each word, and language models giving the probability of word sequences which is normally estimated from the word transcriptions (Goddeau and Zue, 1992; Jelinek, 1990; Kuhn and De Mori, 1995; Zue et al., 1990).

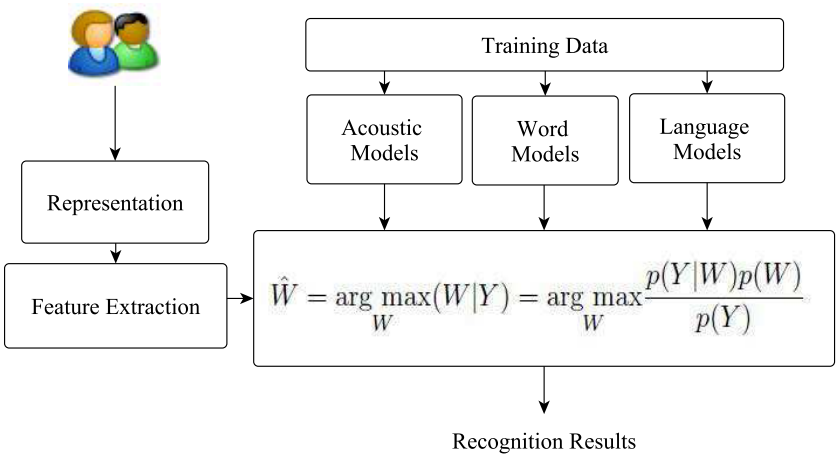



Figure 2.4: The structure of a typical HMM-based speech recognizer.

Given the feature vectors for the speech signal Y, the HMM-base speech recognizer finds the most likely word sequence using maximum a-posteriori estimation,

$$W = P(\text{feature vectors}|W)P(W), \tag{2.21}$$

where $W = W_1, W_2, \ldots$ is the sequence of words. The right hand side has two probabilistic models. The first is the acoustic model. It calculates the probability of the feature vectors given a sequence of words, $P(\textit{feature vectors}|W_1, W_2 \ldots)$. Acoustic models are usually defined on linguistic units, e.g.,phones and words. There exist many types of acoustic models, for example, allophone models, polyphones, and allophones. The second is the language model. It calculates the probability of the sequence of words itself, $P(W_1, W_2 \ldots)$. The most commonly used language model are the N-gram language model. It assumes that the probability of any word depends only on the previous N words in the sequence. For example, a 2-gram or bi-gram language model would compute $P(W_1, W_2 \ldots)$ as,

$$P(W_1, W_2, W_3 \ldots) = P(W_1)P(W_2|W_1)P(W_3|W_2)P(W_4|W_3) \ldots . \tag{2.22}$$

Similarly, a 3-gram model would compute it as

$$P(W_1, W_2, W_3 \ldots) = P(W_1)P(W_2|W_1)P(W_3|W_2, W_1)P(W_4|W_3, W_2) \ldots . \quad (2.23)$$

In practice the language probability is raised to an exponent for recognition, which is frequently referred to as the language weights. Optimal values of $\alpha$ typically lie between 6 and 11.

Fig. 2.5 illustrates the HTK implementation of a typical HMM recognizer. HMM-based speech recognizer consists of two stochastic processes, a hidden Markov chain, which accounts for the temporal variability, and an observation, which accounts for the spectral variability. This combination allows speech models, e.g., phone models and word models, to form a large and complete network to produce the most likely sequence of words or phonemes in ASR (Cole et al., 1997; Lee et al., 1992, 1990).

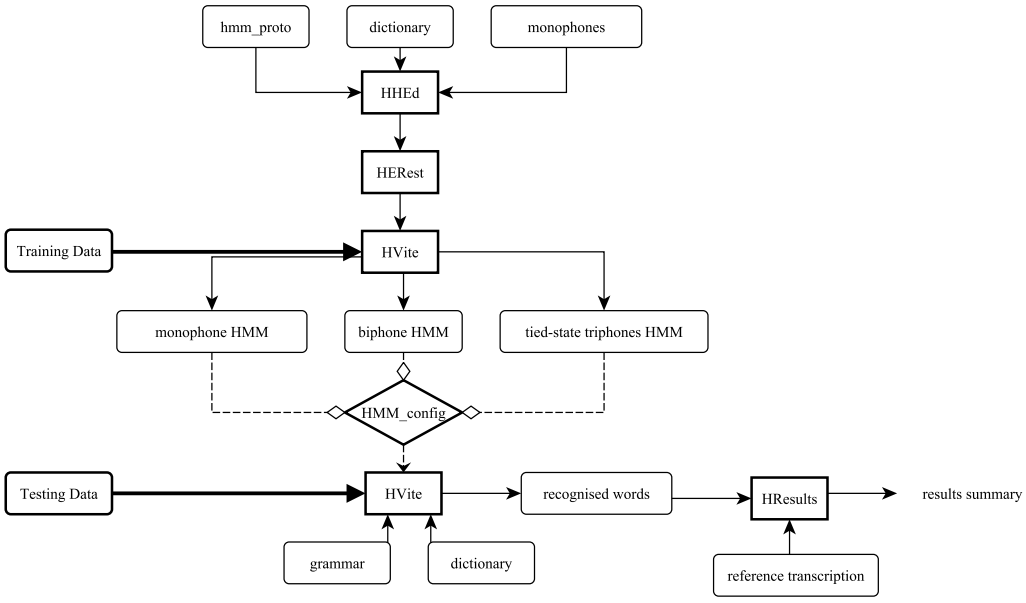



Figure 2.5: Implementation of the HMM-based speech recognizer using HTK.

### 2.4.2 Connectionism

Connectionist model is another useful paradigm in ASR. It uses parallel collections of simple processing elements densely connected by weights whose strengths

are modified through learning and training to mimic the neuron connections in the human brain. The connectionist structure can directly integrate multiple knowledge sources, and provides a distributed form of associative memory, which is known as artificial neural networks (ANN) (Dede and SazlI, 2009; King, 1997; Lang et al., 1990; Lee et al., 1988; Muller et al., 2006).

#### 2.4.2.1 Artificial Neural Network

ANN consists of individual units termed neurons, as shown in Fig. 2.6. Through back-propagations, the weights can be adjusted during training when a particular pattern at the input is observed. The back-propagation algorithm calculates the difference between the desired and actual outputs and modifies the weights proportionally (Waibel and Lee, 1992). Speech signal is essentially a left-to-right time sequence. Several neural network architectures have been developed for ASR development,

1. multi-layer perceptrons (MLPs): it uses an input buffer, e.g., a sliding window, to transform a temporal pattern into a spatial one for pattern matching (Bourlard and Morgan, 1993; Lippmann, 1989),

2. time-delay neural networks (TDNNs): it enables online adaption and prediction of continuous speech data, where output is computed as a function of the previous input in the time series (Poo, 1997; Waibel and Lee, 1992),

3. Recurrent networks (RNNs) : it accepts sequential input vectors using a context layer to retain information between adjacent time frames (Robinson and Fallside, 1991; Schuster and Paliwal, 1997).

#### 2.4.2.2 ANN-based Speech Recognition

Similar to statistical models, connectionist models also require training in ASR. However, unlike HMMs, ANNs do not make assumptions about the underlying probability distributions of the data. It has easier hardware implementations and higher operation efficiency than HMMs (Schwarz et al., 2006). ANNs are applied for various tasks ranging from simple voiced/unvoiced checking to complex

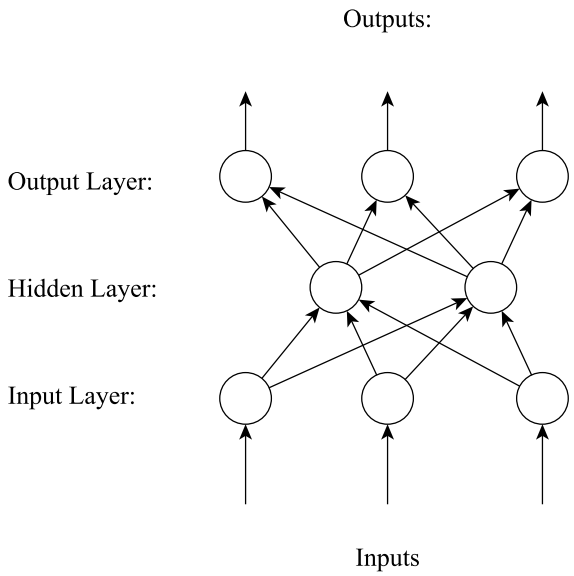



Figure 2.6: A classic feedforword neural network.

phoneme classification and word recognition. In static classifications, the neural network sees all of the input before it makes a decision, e.g., MLPs (King and Taylor, 2000). In dynamic classification, the neural network sees only a small window of the input, e.g., TDNNs, RNNs (Bortman and Aladjem, 2009). The window slides over the input frames while the network makes a series of local decisions, which are integrated to make a global decision. Static ANNs works well for phoneme recognition, but worse on words recognition compared to dynamic ANNs (Bourlard and Morgan, 1993; Lippmann, 1989).

A simple experiment performed by Huang and Lippmann demonstrated the ability of ANNs in ASR. They used MLP with only 2 inputs, 50 hidden units, and 10 outputs, on a collection of vowels produced by men, women, and children. Using the first two formants of the vowels, $F_1$ and $F_2$, as the input feature vectors, the network produced the decision regions in the acoustic domain after $50,000$ iterations of training (Bourlard and Morgan, 1993; Lippmann, 1989). The decision regions resembled that of the human listeners. A more complex network constructed by Elman and Zipser achieved error rates of as low as 0.5% for vowels and 5.0% for consonants (Elman, 1988). Waibel et al. also demonstrated excellent

results for phoneme classification using a TDNN (Waibel and Lee, 1992). The final output is computed by integrating the 9 frames of phoneme activations in the second hidden layer. Their TDNN was trained and tested on 2000 samples of /b/, /d/, and /g/ phonemes manually excised from a database of 5260 Japanese words. It achieved an error rate of 1.5%, compared to 6.5% of their HMM-based recognizer (Waibel and Lee, 1992). The design of time-delay has several desirable qualities,

1. the compact structure economizes on weights and forces the network to develop general feature detectors,

2. the hierarchy of delays optimizes these feature detectors by increasing their scopes at each layer,

3. its temporal integration at the output layer makes the network insensitive to the exact positioning of the speech event.

Schuster and Paliwal (1997) also obtained good phoneme recognition performance using RNNs. In addition, they proposed to integrate two sets of RNNs with time-delay to account for the contextual information in the speech signal, and achieved moderate improvements on the Texas Instrument and Massachusetts Institute of Technology (TIMIT) corpus.

#### 2.4.2.3 Hybrid HMM/ANN Methods

A new family of classifiers combining HMM with ANN has also been developed and implemented for difficult ASR tasks, the hybrid HMM/ANN (Bourlard and Morgan, 1993; Kanazawa et al., 1995; Sim and Bao, 1998; Sirigos et al., 2002). It uses the connectionist structure to model the time-indexed feature vectors generated by the HMM. Each output unit of an ANN is associated with one HMM state. Then ANNs generate the posterior probabilities of the state. This probability can be used as local probabilities in HMMs (Bourlard and Morgan, 1993). The hybrid approach obtains better or equivalent results compared to the HMM system. It is also efficient in terms of CPU usage. It has no strong assumptions about the statistical distribution of the acoustic space, which improves the robustness even with insufficient training data (Lubensky et al., 1994; Renals et al.,

1994; Steeneken and Van Leeuwen, 1995). Detailed review of ANNs in ASR can be found in (Lippmann, 1989; Trentin and Gori, 2001).

## 2.5 Robustness Techniques

Natural speech contains many variations. ASR applications are especially challenging dealing with the background noises and the speaker variations.

### 2.5.1 Noise Contamination

The approaches to improve noise robustness in ASR systems can be grouped into three categories (Flynn and Jones, 2008; Mitra et al., 2011),

1. the front-end based approach,
2. the back-end based approach,
3. the missing feature theory.

The front-end based approaches usually aim to generate relatively descriptive and discriminative features for the back-end classifier. For example, spectral subtraction (Kim et al., 2004), auditory scene analysis (Brown and Cooke, 1994; Rouat, 2008), and auditory modeling (Flynn and Jones, 2008; Jeon and Juang, 2007) have shown good results on noise robustness. The back-end based approach focuses on reducing the mismatch between the training and the testing conditions in the recognizer. For example, different types of noise at different levels are used to train the back-end models (Gong, 1995; Rabaoui et al., 2004; Windmann and Haeb-Umbach, 2009). Adaptation methods can also be used as an alternative to improve noise robustness (Holmberg et al., 2006). Maximum-likelihood linear regression (MLLR) performs model adaptation by rotating the Gaussian mixture means of clean HMMs using linear regression without using any prior knowledge of the background noise (Afify et al., 2009; Mitra et al., 2011; Suh et al., 2007). A modified version of MLLR was also proposed for piecewise-linear transformation, where different noise types are clustered based on their spectral characteristics (Mitra et al., 2011). Separate acoustic models are trained for each noise cluster at

different signal-to-noise ratio (SNR). During recognition, the best matched HMM is selected and adapted by MLLR. The third approach, the missing feature theory, assumes that the speech regions that are contaminated by noise can be treated as missing (Cui and Alwan, 2005; Gemmeke et al., 2010; Huang and Juang, 2003; Raj and Stern, 2005). It computes a time-frequency reliability mask to find reliable regions and the unreliable regions so as to make a binary decision. Once the mask is obtained, the unreliable regions are dealt with by two methods: 1) data imputation where the regions are re-estimated based on the reliable ones, and 2) marginalization where only the reliable regions are used by the back-end recognizer.

### 2.5.2 Speaker Variation

In the absence of noise, the speaker-specific variations, e.g., coarticulations and contextual effects, are the main causes of ASR performance degradation (Cole et al., 1997; Sankar et al., 2001). The physiological characteristics of individuals directly affect the acoustic phonetic qualities (Sankar et al., 2001; Weenink, 2006). Speaker differences are usually dealt with by adapting the acoustic model to a particular speaker. A simple method is to normalize the vocal tract length (Claes et al., 1998). In addition, contextual variations can by accounted for by tri-phone or bi-phone models. These models represent speech as a sequence of non-overlapping phonetic units. However, they often suffer from data sparsity and may only capture the contextual influence from the immediate neighboring phones. For instance, coarticulation can have contextual influences beyond the immediate neighbors. In this direction, the speech production knowledge is promising. For instance, the articulatory attributes of phones are slow varying and much less variant compared with the acoustic attributes (Wrench and Richmond, 2000).

### 2.5.3 Articulatory Cues

In recent years, many articulatory based processing methods have been proposed separately to address the problem of pronunciation variations in a number of frame-based, segment-based, and acoustic landmark systems (King et al., 2007;

Scharenborg et al., 2007; Stevens, 2002). Real life articulatory data such as electromagnetic articulograph (EMA), X-ray analysis, and laryngograph provide good references for physiological speech studies, but the current techniques are still invasive and often unrealistic for on-line ASR operations (Richmond, 2009). Manually derived phonological articulatory features (PAFs) from phonological rules, e.g., manner of articulation (MOA) and place of articulation (POA), have also been used in ASR applications (King et al., 2007; Kirchhoff et al., 2002; Saenko et al., 2005). Speech synthesis models have also been used to search the acoustic correlates for the highly error-prone phoneme classes in ASR, such as stops (Blumstein et al., 1977), fricatives (Heinz and Stevens, 1961), and nasals (Malecot, 1973; Recasens, 1983). Initial research attempted to incorporate speech production knowledge by deciphering appropriate features to capture articulatory dynamics. Deng et al. used 18 AFs to describe the place of constriction, horizontal and vertical tongue body movements, and voicing information (Deng et al., 2004). They reported an average classification improvement of 26% over the conventional HMM system for a speaker-independent task. Phone recognition on the TIMIT dataset showed a relative improvement of about 9% over the MFCC-HMM baseline.

Using articulatory trajectory information is challenging because it needs to retrieve the articulatory dynamics from the speech signal, which is known as the acoustic-to-articulatory mapping or the speech inversion (Behbood et al., 2011; Schroeter and Sondhi, 1994; Youssef et al., 2009). One of the earliest works on acoustic-to-articulatory inversion used temporal decomposition to predict the corresponding vocal tract configuration from acoustic signal (Atal and Hanauer, 1971). Ladefoged et al. used MLPs to estimate the shape of the tongue in the midsagittal plane, using the first three formant frequencies in consonant vowel (CV) sequences. Codebook-based approaches have also been proposed for the inversion task. Richmond proposed mixture density networks (MDNs) to obtain flesh-point trajectories, the pellet trajectories, as conditional probability densities of the input acoustic parameters. Compared with MLPs, MDNs more directly address the non-uniqueness issue (Mitra et al., 2011; Richmond, 2009). However, inversion studies have mostly been confined to predicting the dynamics. An alternative is to use the articulatory recordings. Frankel et al. modeled the recorded

articulatory trajectories using phone-specific linear dynamic models. The directly measured articulatory data in conjunction with MFCCs showed a performance improvement of 9% over the MFCCs baseline.

Many authors mapped phonemes to the canonical PAFs based on the phonological properties of speech sounds (Kirchhoff et al., 2002; Siniscalchi and Lee, 2009), as shown in Table 2.1. The value *nil* indicates non-applicable categorization. The APFs in this study shares a common ground with PAFs in that they both embed production knowledge in ASR. However, the pronunciation modeling method in this study distinguishes the base-forms from the variations of English phonemes. A more reliable heuristic learning strategy is also designed to retrieve the APFs for speech recognition experiments. Generally speaking, both PAFs and APFs are alternatives of directly using recorded physiological data. However, there are a few notable differences. The proposed APFs are continuous valued representations, which are closer to actual articulatory realizations than the manually-mapped discrete-valued PAFs. PAFs have by far taken a dominant position in ASR since they require no additional computations or invasive recording devices. The mapping between the canonical PAFs and the 61 TIMIT phonetic base form annotations is straight forward (Frankel et al., 2007; King and Taylor, 2000; Kirchhoff et al., 2002). For example, the mapping of phones using the MOA attribute are shown in Fig. 2.7, POA in Fig. 2.8, and voicing in Fig. 2.9.

The following acoustic parameters are used as the mapping criterion as introduced in (Ali et al., 2006):

- $F_i$: $i^{th}$ formant;
- $VF_i$: $i^{th}$ formant of the neighboring vowel in CV patterns;
- VOT: Voicing onset time;
- BF: Burst frequency;

Normally $F_1$ and $F_2$ are sufficient to map the English vowels and some of the plosives. $F_3$ would provide some further distinction between formants regions. Moreover, VOT is known to determine whether a particular stop consonant is

Table 2.1: PAFs derived from English phonological rules.

| PAF | Value |
|---|---|
| MOA | silence, plosive, fricative, affricate, nasal, approximant, flap, vowel |
| POA | silence, bilabial, labiodental, dental, alveolar, post-alveolar, palatal, velar, glottal, glide, diphthong |
| Voicing | silence, voiced, unvoiced |
| Height | silence, nil, open, middle, close, glide, diphthong |
| Frontness | silence, nil, front, central, back, glide, diphthong |
| Rounding | silence, nil, unrounded, rounded, glide, diphthong |

perceived as voiced or voiceless regardless of the place of articulation Halliday and Webster (2006). BF is defined as the most prominentfrequency during the voicing release,

$$BF = min_j(argmax_i S_{ij}), \tag{2.24}$$

where $j$ is the time during burst, $i$ is the number of filters, and $S$ is the spectral envelope. The formant transition (FT) is calculated as,

$$FT_i = \frac{max_j(F_{ij} - F_{(i-1)j})}{max_j(F_{ij} + F_{(i-1)j})}. \tag{2.25}$$

The thresholds, $FT\_TH$ and $VF\_TH$, in the MOA and POA maps are empirically defined (Ali et al., 2006).

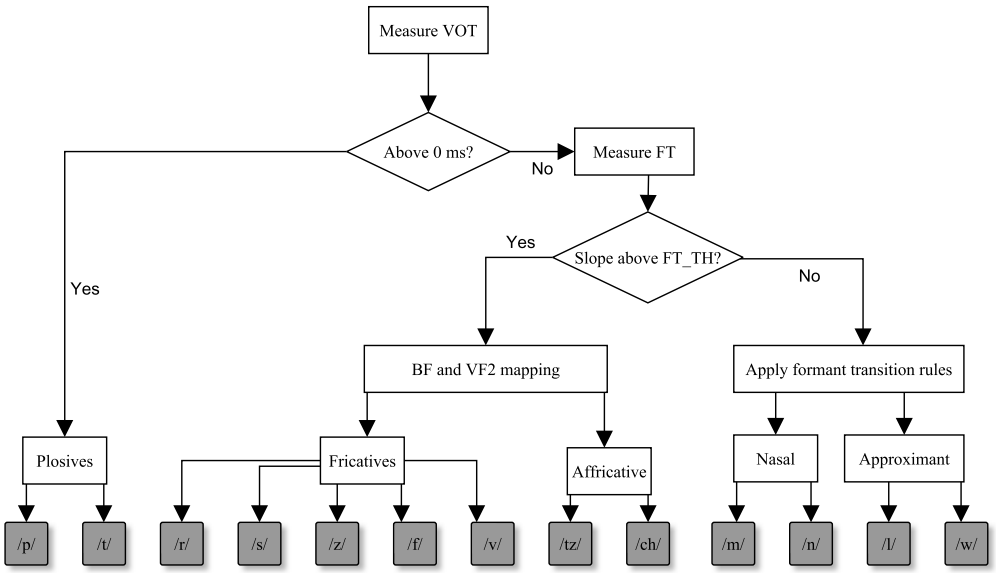



Figure 2.7: Illustration of the mapping between the MOA features and the English phonetic base forms.

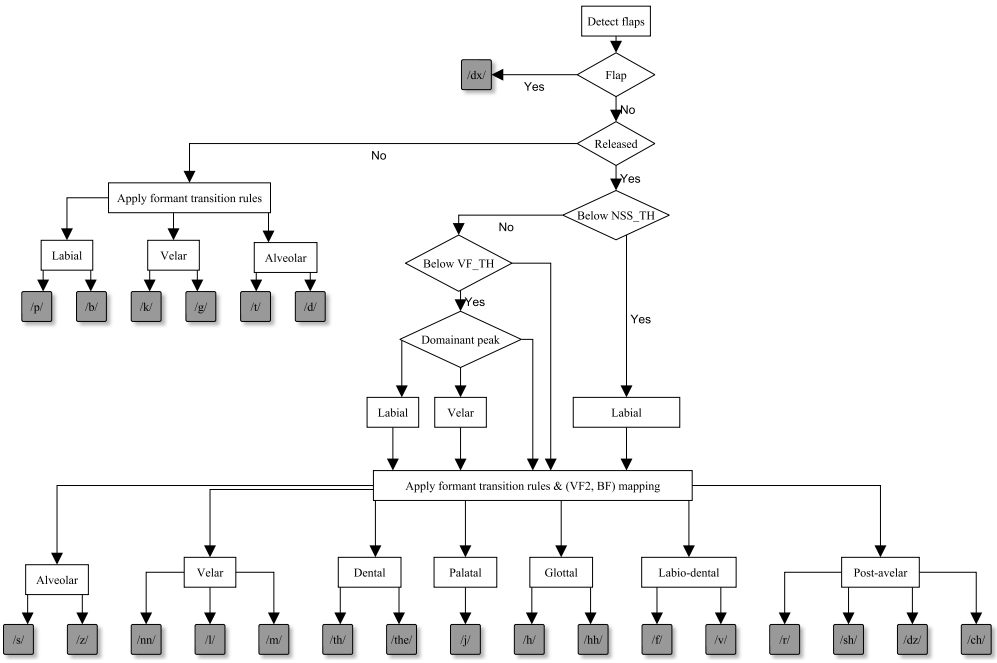



Figure 2.8: Illustration of the mapping between the POA features and the English phonetic base forms.

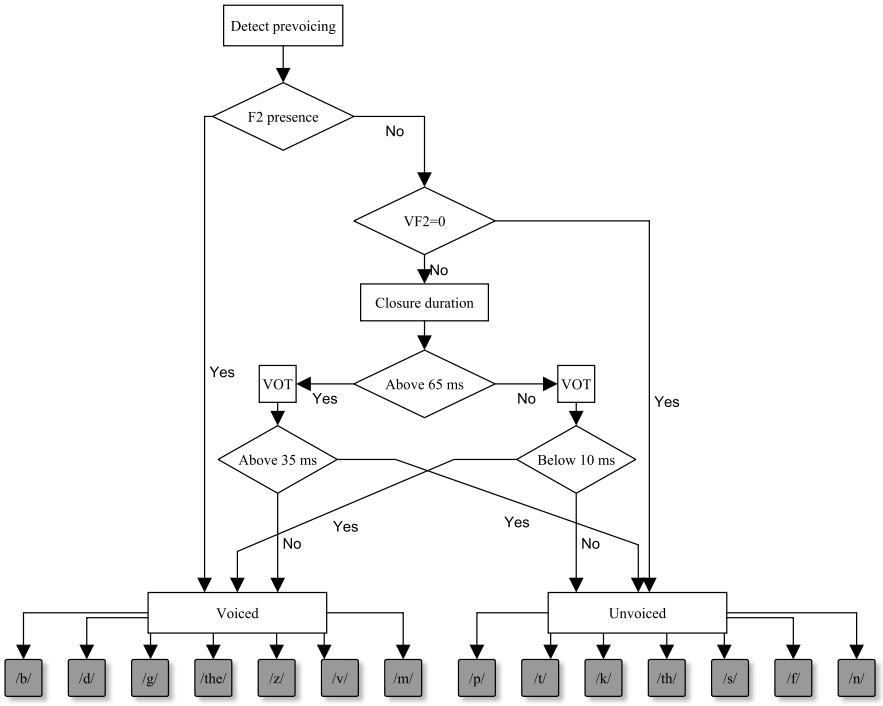



Figure 2.9: Illustration of the mapping between the voicing features and the English phonetic base forms.

# Chapter 3

# Adaptive Neural Control Scheme for Articulatory Synthesis

## 3.1 Overview

Reproducing the smooth vocal tract trajectories in continuous speech is critical for high quality articulatory speech synthesis. This chapter presents an adaptive neural control scheme based on fuzzy logic and neural networks [1]. The proposed controller uses the controller to track the articulatory movements of the human vocal tract and infer the activation patterns of the underlying muscular structures. It achieves high accuracy during on-line tracking of the lips, the tongue, and the jaw in the simulation of consonant-vowel sequences. Furthermore, the controller manipulates the mass-spring based elastic tract walls in a 2-D articulatory synthesizer to realize efficient speech motor control. The neuron controller serves to construct the articulatory-acoustic mapping of English phonemes. It also offers salient qualities such as generality and adaptability for future developments of control models in articulatory synthesis.

The chapter is organized as follows. Section 3.3 formulates the control of the articulatory dynamics in the mass-spring based 2-D vocal tract system. Section 3.4 describes the structure, the learning algorithm, and the adaptive laws of the proposed extended fuzzy neural network (E-FNN) controller. Section 3.5 reports

[1]The original manuscript of this chapter was submitted to Computer Speech and Language in July 2012.

the experimental settings and the simulation results. Section 3.6 discusses the present and future aspects of speech motor control.. Section 3.7 summarizes the chapter.

## 3.2 Control of Articulatory Dynamics

Concatenative synthesis is a popular method in text-to-speech (TTS) applications. It uses the stored speech waveforms of phonemes or words pronounced by the human speakers to generate intelligible output. However, its applications are limited to the available speech and speakers. In contrast, articulatory synthesis simulates the movements of the speech apparatus for TTS applications. It has a stronger physiological basis and is able to produce a larger number of utterances than the concatenative method. Moreover, in applications such as the facial animation (Badin et al., 2002), the medical treatment of speech disorders (Kröger et al., 2008), and the articulatory-phonetic studies in automatic speech recognition (ASR) (King et al., 2007), the articulatory synthesis method offers additional benefits beyond TTS. For example, Mitra et al. (2011) used the articulatory synthesizer to prepare a codebook of acoustic-to-articulatory data pairs, and they showed that the inferred articulatory features increased the robustness toward noise contamination and speaker variations in the ASR systems.

A complete articulatory synthesizer usually includes three functional components: an anatomical model, an acoustic model, and a control model. Studies on the anatomical and the acoustic models have developed rapidly in the past decades (Birkholz et al., 2007; Buchaillard et al., 2009; Cook, 1990). However, it remains a challenging task to find an efficient control strategy in current articulatory synthesis research. For instance, the control model should be able to reproduce realistic articulatory trajectories in different phonetic contexts and even with different speaking rate. Existing models often operate in a codebook fashion, which applies a set of manually derived linguistic rules to define the articulatory targets such as the velocity or the position profile of a particular speech sound e.g., a phone. This kind of synthesis-by-rule approach was initially implemented in Ishizaka & Flanagan's cord-tract model (Flanagan et al., 1975) and Saltzman's task-dynamic articulatory model (Saltzman and Munhall, 1989).

Usually each phone has one spatial target in the codebook. The articulatory movements for the sequential phonetic strings such as syllables, words, and sentences, are generated by interpolating and/or approximating the targets (Birkholz et al., 2011; Perrier et al., 2005). Different from the codebook approach, Nelson (1983) suggested that the articulatory movements were the result of optimized control similar to that of a second-order dynamical system. Löfqvist and Gracco (2002) supported the view, and they observed that a cost minimization principle could well explain the trajectory curvature of the articulatory kinematics. Generally in speech motor control, the equation of motion that governs the dynamics of the articulators is analogous to a mass-spring damper (MSD) (Kelso et al., 1986; Kröger et al., 1995; Perrier and Ostry, 1996), which follows Newton's law,

$$u + F_e + F_f = M\ddot{z} + B\dot{z} + Kz, \tag{3.1}$$

where M, B, and K are the mass, damping, and stiffness coefficients of the speech articulators, e.g. the jaw, tongue tip, and lower lip in the anatomical model. $u$ is the input force or the activation level of the muscular structures which control the TVs. $F_e$ is the external force due to the gravity factor and the air pressure inside the tract, and $F_f$ is the friction force between the adjacent muscular structures, which can be assumed to be negligible due to the saliva. We can describe the articulators using the vocal tract variables (TVs) (Saltzman and Munhall, 1989), where $z, \dot{z}$, and $\ddot{z}$ are the position, the velocity, and the acceleration parameters, respectively. For example, the masseter controls the jaw movements. We refer to the set of muscular activation forces $u$ as the motor variables (MVs).

The equation of motion describes the quasi-incompressibility of the speech articulators during the speech production. It yields close-loop solutions by choosing the appropriate time-variant variables. For example, Feldman's equilibrium point hypothesis (EPH) used the equilibrium positions as the time-variant variables, the shift of which results in the movements of the articulators (Feldman, 1986). Perrier et al. and Buchaillard et al. applied the EPH control concept in a finite element model of the tongue, and solved the differential equation using combined Newton-Raphson and Newmark method (Buchaillard et al., 2009; Perrier and Ostry, 1996; Perrier et al., 2003). In contrast, Saltzman and Munhall (1989) deemed the stiffness coefficients as the time-varying variables. They used

a pseudo-Jacobian inversion matrix to calculate the gestural control parameters for the desired articulatory movements in the differential equation. Still their concept resembles the EPH method, since the stiffness directly affects the velocity with which the equilibrium length is restored (Boersma, 1998). The concept is also used in the articulatory synthesizers in (Birkholz, 2005; Kröger et al., 2009, 1995). However, the method requires explicitly defining the gestural scores and the activation intervals for the TVs in the control model, which are highly error prone especially at the phonetic boundaries. Another option is to use the time-varying input force functions to reproduce the system state profile $[\dot{z}, z]$, the velocity and position trajectories (Kröger et al., 1995). Here the coefficients M, K, and B assume values that are close to human tissues in the vocal tract. In this direction, the equation of motion (3.1) simplifies to an ordinary differential equation. For example, van den Doel and Ascher (2008) formulated a wall displacement model,

$$M\ddot{z}(x,t) + B\dot{z}(x,t) + K\vec{z}(x,t) = p(x,t), \tag{3.2}$$

where the driving force is the air pressure p inside the tube. Various discretization techniques such as the leap-frog scheme in (Boersma, 1998) and the Newmark methods in (van den Doel and Ascher, 2008) are then used to obtain the solution during the articulatory and acoustic simulation. One drawback is the high computation cost which is inefficient for on-line articulatory control.

The dynamic MSD system is highly non-linear and contains uncertainties which are difficult to describe with precise mathematical model. First, human vocal system consists of soft tissues as well as bony structures e.g., the hard palate. Consequently, the MSD system contains unmodeled variabilities in the M, B, and K parameters, which vary from speaker to speaker (i.e., physiological differences) and for the same speaker under different conditions (e.g., emotional states). Moreover, the stiffness of muscular tissues changes during activation (Duck, 1990; Perrier et al., 2003). Second, during speech production, the modeling of constriction is not linear. Though the articulatory movements are smooth between vowel targets, the transitions to/from the consonants such as plosives, nasals, and laterals, are not so. For example, when the tongue tip hits the alveolar ridge during [d/t] production, the collision introduces points of discontinuity

in the vocal tract at the onset of the closure, rendering the model non-linear (Birkholz et al., 2011). Third, the articulatory movements are also affected by the phonetic structure of continuous speech, which introduce unmodeled variabilities. Therefore, there is an urgent need for more adequate dynamic modeling methods to deal with the above issues and to realize efficient speech motor control.

Artificial neural networks (ANNs) have shown advantages in non-linear modeling of dynamic control systems. For example, Saltzman and Munhall proposed to use Jordan's recurrent networks (RNN) to incorporate the temporal dynamics and learning algorithm in the control model (Jordan, 1986; Saltzman and Munhall, 1989). Hirayama et al. (1993) applied ANNs to learn the inverse dynamics of speech motor control. More recently Fang et al. used a general regression neural model to infer motor commands from the articulatory measurements (2009). However, the fix-structured ANNs usually use a trial-by-error approach to determine the parameter and structure in the neural controller. As a result, the controller performance is subject to the experimenter's decision rather than the property of the dynamic system. In this aspect, ANNs with fuzzy logic, or fuzzy neural networks (FNNs) are more appropriate than the fix-structured ANNs (Wang, 1997). For example, Kröger et al. (2009) used self-organizing maps to learn the motor commands and the tract variables on the phonetic sequences, which obtained encouraging results in the articulatory synthesizer.

Previously an adaptive neural controller have been introduced, termed the generalized dynamic fuzzy neural network (GD-FNN) controller (Er and Gao, 2003; Wu et al., 2001). The controller has shown excellent performance in terms of tracking accuracy and computational efficiency for several non-linear dynamic systems with uncertainties, e.g., an inverted pendulum, a robot manipulator (Gao and Er, 2003), and a drug delivery system (Gao and Er, 2005). In this study, the adaptive neural control model is applied to reproduce the articulatory trajectories of the vocal apparatus in a 2-dimensional (2-D) articulatory synthesizer. The GD-FNN can infer knowledge about the articulatory dynamics and stores the information in the neural structures and the fuzzy logics. The control scheme presented here is based on an extended version of GD-FNN, referred to as E-FNN. The E-FNN integrates the radial basis function-based neural networks (RBF-NNs), the fuzzy inference network, and the RNN in one neural topology. The

recurrent layer is added to the original GD-FNN to deal with the temporal dynamics of the articulatory speech patterns (Jordan, 1986). The E-FNN controller also embeds a learning algorithm and an adaptive control law to simultaneously determine the structure and the parameters of the neural topology. The purpose is to investigate the mapping characteristics of the muscle activities, the MVs, with the articulatory trajectories, the TVs, in speech motor control. Unlike the TVs, the MVs are usually hard to measure or not completely retrievable in human speech production. The E-FNN learns to predict the MVs from the TVs using the generalization abilities of fuzzy logics and ANNs. The E-FNN model is then coupled with a proportional integral derivative (PID) controller to manipulate a MSD system to reproduce the continuous and smooth articulatory trajectories of the desired consonant vowel (CV) sequences. The tracking accuracy of the E-FNN controller is reported in comparison with the electromagnetic articulograph (EMA) data of the vocal tract in CV articulation.

## 3.3 Articulatory Dynamics

The controllability canonical form for a $2^{nd}$ order time-variant non-linear system is (Slotine and Li, 1991),

$$\ddot{z}(t_s) = f_n(\underline{z}, t_s) + g_n(\underline{z}, t_s)u(t_s) + d_n(t_s), \tag{3.3}$$

where $\underline{z} = [\dot{z}, z]$ is the state vector, velocity and position, of the system, $f_n$ and $g_n$ represent the non-linearities of the mapping from the input u to the output z, and d represents the uncertainties and external disturbances of the dynamic system, and $d_n$ is the unmodeled uncertainties. If we further define the non-linear dynamic function $f_n(\underline{z}, t_s) = f(\underline{z}, t_s) + \Delta f(\underline{z}, t_s)$, and the control gain $g_n(\underline{z}, t_s) = g + \Delta g(\underline{z}, t_s)$, where $f$ and $g$ are the nominal parts, $\Delta f$ and $\Delta g$ are the unknown part or the uncertainties of f and g (Lin and Li, 2012), the canonical form can be re-written as,

$$\ddot{z}(t_s) = f(\underline{z}, t_s) + gu(t_s) + d(t_s), \tag{3.4}$$

where $d(t_s) = \Delta f(\underline{z}, t_s) + \Delta g(\underline{z}, t_s) + d_n(t_s)$ is the unknown uncertainties From the equation of motion of the MSD system in (3.1) and the controllability canonical

form in (3.4), we have

$$f = -\frac{B}{M}\dot{z}(t) - \frac{K}{M}z(t), \tag{3.5}$$

and

$$g = \frac{1}{M}. \tag{3.6}$$

Since $g \neq 0$ for all $\vec{z}$, the system is controllable (Gao and Er, 2003; Lin and Li, 2012; Slotine and Li, 1991). Furthermore, the function $f$ and $d$ are assumed to be bounded in human vocal system.

This study focuses on the control of the vocal tract including the lips, the tongue, and the jaw in a 2-D articulatory synthesizer, as shown in Fig. 3.1, which was constructed by Mermelstein based on the X-ray image of a human speaker (Mermelstein, 1973). In particular, the vocal tract has elastic walls analogous to the MSD. The 12 TVs are the pallet points in Fig. 3.1, including the x-y coordinates of the tongue root (TR) $(TR_x, TR_y)$ relative from its neutral or resting position, the tongue body (TB) $(TB_x, TB_y)$, the tongue tip (TT) $(TT_x, TT_y)$, the lower lip (LL) $(LL_x, LL_y)$, the upper lip (UL) $(UL_x, UL_y)$, and the lower incisor (LI) $(LI_x, LI_y)$. The elastic mass-springs are used to represent the MVs, which underly the TVs in the 2-D vocal tract. The 8 MVs include one intrinsic tongue muscle: superior longitudinal (SL), which retracts or flaps the tongue tip; four extrinsic tongue muscles: anterior genioglossus (GGa), posterior genioglossus (GGp), hyoglossus (HG) and styloglossus (SG), which change the shape and position the tongue dorsum: body and root; three facial muscles: masseter (MA) which raises or lowers the jaw, risorius (RO) and orbicularisoris (OO) which constrict, round, or spread the lips. The vocal cords, shown as the glottis in Fig. 3.1, are not included in this model for two reasons. First, the control of vocal cords is more effective with stiffness parameters than the MV parameters (Flanagan et al., 1975). Second, the vocal cords can cause non-unique mapping between the TVs and the MVs by compensating the vocal tract change in speech production (Schroeter and Sondhi, 1994). For example, if we are to model the simple voicing contrast for the [p/b] and the [t/d] pairs, additional control variables regarding the timing of glottal excitation need to be specified in the vocal cords.

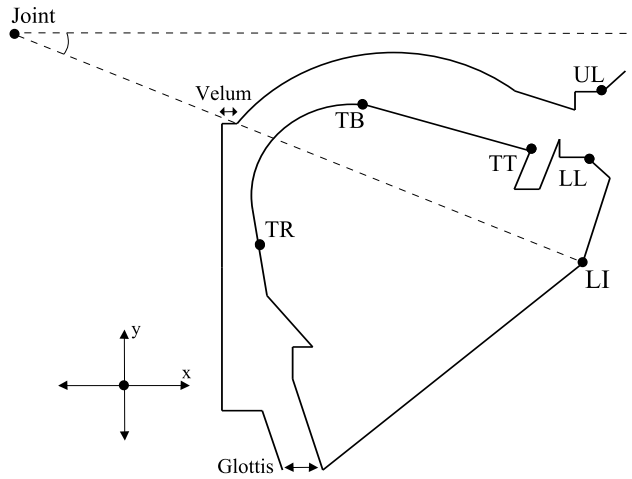



Figure 3.1: Illustration of Mermelstein's 2-D articulatory mesh and the location of vocal tract variables.

## 3.4 Neural Control Scheme

As shown in Fig. 3.2 (a), during off-line training, the proposed E-FNN model learns the inverse characteristics between the input MV, $u$, and the output TV, $z$, in the dynamic MSD system. Since the exact MVs are unknown, and the parameters $M$, $B$ and $K$ vary from speaker to speaker and for the same speaker in different phonetic contexts, the E-FNN controller uses the reference MVs, $u_r$, (details given in Section 3.5) and the desired TVs, $z_d$, to learn the system non-linearities and dynamics through an embedded learning algorithm. Using the training data pairs, the algorithm determines the structure and the parameters of the E-FNN, e.g., the number of hidden neurons and the weights systematically and automatically through an iterative supervised learning (3.4.2). During on-line tracking, as shown in Fig. 3.2 (b), instead of looking for exact MVs, the PID controller generates the compensation output and the tracking error at each sample time for the overall control system. It embeds an adaptive control law, which uses the error rate as the weight update criteria and stores the system dynamics and the mapping functions in the E-FNN (3.4.3). In this manner, the E-FNN controller infers the muscular activation patterns by tracking the articulatory movements.

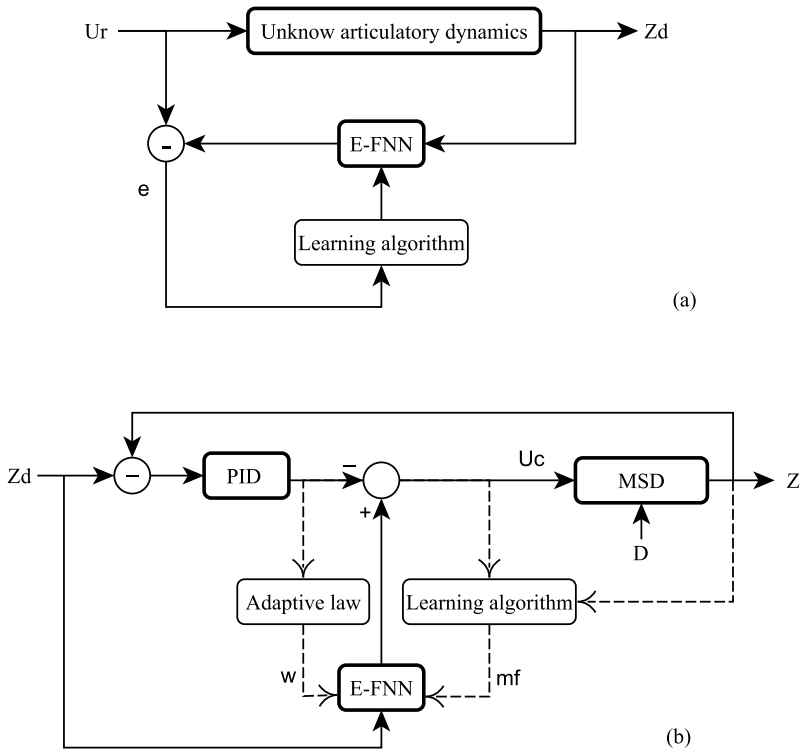



Figure 3.2: Structure and data flow in the proposed fuzzy neural controller. (a) Off-line learning on the training data pairs; (b) On-line tracking of the desired articulatory trajectories in the MSD based vocal tract system.

### 3.4.1 E-FNN Structure

The E-FNN architecture is shown in Fig. 3.3, which has a total of five layers. It incorporates the Takagi-Suegeno-Kang-type fuzzy inference system, the RBF-NN, and the RNN in a connectionist structure, which is extended from the GD-FNN (Er and Gao, 2003; Gao and Er, 2005; Wu et al., 2001). The recurrent layer accounts for the temporal dynamics in speech motor control (Jordan, 1986). Nodes and links in layer one and two act as a fuzzifier, while nodes and links in layer four act as a defuzzifier. Here $x_i^{(l)}$ denotes the $i^{th}$ input of a node in the $l^{th}$ layer and $y^{(l)}$ denotes the node output in layer l. The function of the node in each layer is given in the following.

- Layer 1: the input linguistic layer. Each node transmits the input variable

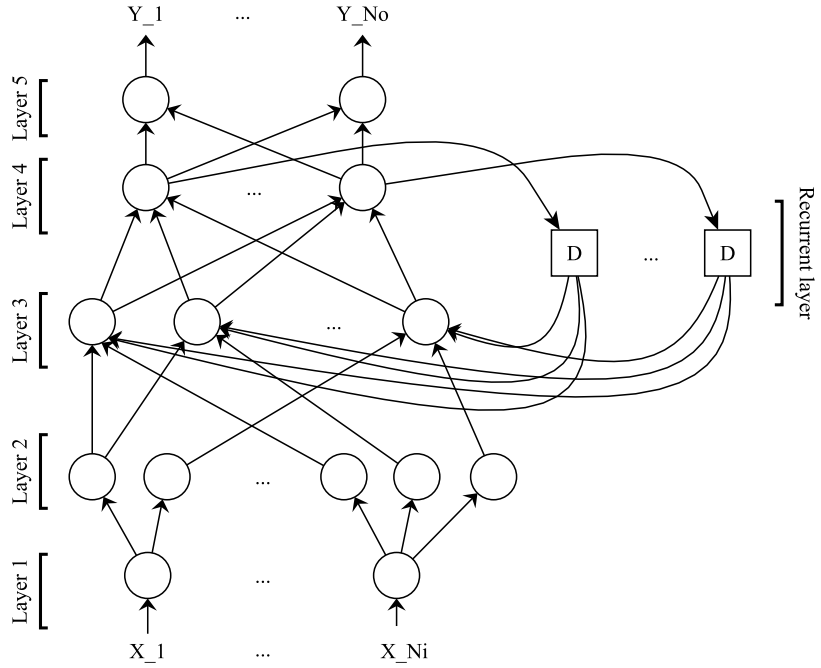



Figure 3.3: Architecture of the extended fuzzy neural network.

to the next layer directly,

$$y_i^{(1)} = x_i^{(1)}, i = 1, \cdots, N_i. \tag{3.7}$$

For the inverse control model, $N_i = 36$, which includes the position, velocity and acceleration of the 12 TVs:$[\ddot{z}, \dot{z}, z]$ .

- Layer 2: the membership function layer. It specifies the degree to which an input variable belongs to a fuzzy set using Gaussian membership function,

$$y_i^{(2)} = exp\{-\frac{(x_i - c_{ij})^2}{\sigma_{ij}^2}\}, \tag{3.8}$$

where $c_{ij}$ and $\sigma_{ij}$, $i = 1, \cdots, N_i, j = 1, \cdots, N_j$, are the center and the width of the Gaussian function for the $j^{th}$ term in the $i^{th}$ input variable. These parameters are obtained in the learning procedure.

- Layer 3: the rule layer. The number of nodes indicates the number of fuzzy rules. The output of a rule node indicates the firing strength of its

corresponding rule, defined as,

$$y_j^{(3)} = y^{(6)} \prod_{i=1}^{N_i} x_i^{(3)}, \tag{3.9}$$

where $y^{(6)}$ is the output of the recurrent layer.

- Layer 4: the weight layer. The TSK-type fuzzy output weights are obtained in the structure learning procedure. The node output $y_k^{(4)}$: $k = 1, \cdots, N_k$ is the weighted sum of the incoming signals, which is a fuzzy OR operation,

$$y_k^{(4)} = \sum_{k=1}^{N_k} x_k^{(4)} = \sum_{j=1}^{N_j} y_j^{(3)} w_{jk}, \tag{3.10}$$

which integrates the fired rules on the same consequence neuron. The weight is,

$$w_{jk} = K_0 + \sum_{i=1}^{N_i} K_i x_i, \tag{3.11}$$

where K's are manually-set and real-valued parameters.

- Layer 5: the defuzzification layer. Each node in this layer corresponds to one output variable. The output function is defined as

$$y_o^{(5)} = \frac{\sum_{k=1}^{N_k} x_k^{(5)} w_{ok}}{\sum_{k=1}^{N_k} x_k^{(5)}}, \tag{3.12}$$

where $x_k^{(5)} = y_k^{(4)}$, $w_{ok}$ is the link weight from the $k^{th}$ term in layer four to the $o^{th}$ output variable in layer five, $o = 1, \cdots, N_o$, and $N_o = N_k$. In the control model, $N_o = 8$, which is the number of the MVs in the dynamic MSD system. The neural function generates a output in [0, 1], which is the normalized activation level of the MVs.

- Recurrent Layer: it calculates the firing strength of the recurrent variable $r_k = y_k^{(4)}$ to the rule layer. The number of recurrent nodes is the same at that of the output node in layer four. The node acts as a delay line to account for the contextual information in the temporal patterns. The node function is defined as,

$$y_k^{(6)} = \frac{1}{1 + e^{-r_k}}. \tag{3.13}$$

The function can be interpreted as a global membership function, which *remembers* the history of discourse in the recurrent variables (Jordan, 1986). The recurrent outputs are fed back to the rules nodes in layer three, which stores the firing history of the fuzzy rules.

### 3.4.2 Learning Algorithm

The learning algorithm enables simultaneous learning of the E-FNN structure and parameter, which was proposed and implemented in previous studies (Er and Gao, 2003; Gao and Er, 2003; Wu et al., 2001). Structure learning determines the number of membership functions in layer two and the number of fuzzy logic rules in layer three. Parameter learning determines the Gaussian parameters in layer two and the link weights in layer four, i.e., the membership function $y_i^{(2)}(c_{ij}, \sigma_{ij})$, and the weight parameters $w_{jk}$. It uses the semi-closed fuzzy set for membership learning and the linear least square method for weight learning. Structure learning automatically creates or deletes fuzzy rules according to the system error and the error reduction ratio in the E-FNN controller. Learning repeats for each input and output data-pair. The parameters and the structure of the E-FNN are tuned automatically on the training data. Initially there are no fuzzy rules in layer three, and they are created or deleted automatically as the learning proceeds. Detailed mathematical descriptions of the learning algorithm, convergence analysis, and stability analysis of the FNN based controller in dynamic modeling are given in (Gao and Er, 2003).

### 3.4.3 Adaptive Control Law

After obtaining the initial value of the weight vector $w_{ij}$ during the learning process, the E-FNN based controller embeds an adaptive control law to adjust the vector to compensate for the modeling errors in the learning algorithm (Gao and Er, 2005). In this study, the E-FNN controller is connected with a PID controller via adaptive control, as shown in Fig. 3.3 (b). The PID controller serves as a feedback compensator which also stabilizes the inverse dynamic modeling (Gao

and Er, 2005; Lin and Li, 2012). The adaptive control law is designed as follows,

$$u_c(t_s) = u_{E-FNN}(z_d, t_s) - u_{PID}(t_s). \tag{3.14}$$

The PID control output is given by,

$$u_{PID} = K_p e(t) + K_i \int e(t)dt + K_d \dot{e}(t), \tag{3.15}$$

where e is the tracking error, $e(t) = z_d - z$ between the desired target position and the displacement of the MSD. The matrix $K = [K_p, K_i, K_d]$ contains real numbers, and the proper choice of K affects the convergence speed of the tracking performance. The adaptive law adjusts the weight vectors in layer three and four of the E-FNN to minimize the square error E between the desired target position and the estimated position,

$$E(t_s) = \frac{1}{2} u_{PID}^2. \tag{3.16}$$

The discrete gradient method is used to minimize E. The adaptive law of the weight vector is derived as (Wu et al., 2001),

$$\Delta_W = -\eta \frac{\delta E}{\delta W} \tag{3.17}$$

$$= -\eta \frac{\delta E}{\delta u_{S-FNN}} \frac{\delta u_{S-FNN}}{\delta W} \tag{3.18}$$

$$= -\eta \frac{\delta \frac{1}{2}(u_c(t_s) - u_{S-FNN}(t_s))}{\delta u_{S-FNN}} \frac{\delta u_{S-FNN}}{\delta W} \tag{3.19}$$

$$= \eta u_{PID}(t_s) \phi(z_d, t_s) \tag{3.20}$$

where $\eta > 0$ is the learning rate, a small positive number. In neural network training, the learning rate determines the speed at which the system converges. Usually a high learning rate implies high training speed and subsequently better system performance, yet it may also make the system unstable due to weight divergence. In other words, there is a tradeoff between the convergence and stability conditions (Song et al., 2008). For the discrete gradient method used here, the learning rate is determined empirically, which is set as 0.005 in the controller (cf. Section 3.5.3, Chapter 3). In the adaptive controller, the parameter is tuned through trial and error to enable efficient online tracking of the articulatory trajectories.

## 3.5 Simulation

Simulation includes two stages: off-line learning and on-line tracking. In the first stage, the learning algorithm decides the initial weight parameter and the fuzzy rules of the E-FNN topology. It models the inverse dynamics between the motor commands and the tract variables. For this stage, we need to train the E-FNN on parallel MV and TV data. Ideally the training data consists of MVs and TVs measured on the human speech apparatus, such as the electromyography (EMG) and the EMA recordings. EMG measures the motor control of the muscular structures, while EMA measures the corresponding articulatory movements.

### 3.5.1 Data Preparation

This study extracts the CV sequences from the multichannel articulatory (MOCHA) database, which consists of two speakers: one male (MSAK0) and one female (FSEW0), each uttering 460 TIMIT sentences (Wrench, 1999). 807 CV syllables are available in the training data. Each CV sequence has a syllable initial plosive (with or without stress) for every combination of the vowels [ɑ, i, e, ɒ, u] and the plosives [p/b, t/d, k/g] in the pilot study. The EMA data in MOCHA records the movements of the articulators, or the 12 TVs. Similar to Mermelstein's 2-D model, the bridge of the nose and the upper incisor are the reference point in the x-y coordinates (Browman and Goldstein, 1992). The trajectory vectors are z-normalized to have zero mean and unit variance, similar to (Richmond, 2009). The EMA data have at a sampling rate of 500 Hz.

However, reliable EMG data can be difficult to obtain in articulatory studies, e.g., through needle insertion (Baer et al., 1988). Baer observed that the tongue muscles, $GG_a$, $GG_p$, SG, and HG have distinctive level of EMG activation for the cardinal vowels in [ɜpVp] sequences. For example, a single threshold of EMG level can distinguish a front vowel from a back vowel, and each vowel group has consistent activation patterns (Baer et al., 1988). The claim was supported by Buchaillard et al. in the modeling of tongue muscles for cardinal vowel production (Buchaillard et al., 2009). To this end, an alternative set of reference MVs are derived from linguistic studies and the existing EMG recordings to initialize the learning process. As shown in Table 3.1, each phone represents a cognitive

linguistic unit. It corresponds to the motor activity of the muscles, which are normalized to the $[0, 1]$ interval to suit the NN input. For example, GGa is activated (=1) for the front vowel [ɑ], GGp for the high vowels [i] and [u], SG for the back vowels [ɒ] and [u], and HG for the low vowel [ɑ]. To reduce the variability of jaw positions, the MA activation is higher for the low vowel [ɑ] and [ɒ] than for the high vowels [i] and [u]. For the plosive pairs, OO and RO is activated for the labial [p/b], SL and SG for the alveolar [t/d], GGa and GGp for the velar [k/g].

Table 3.1: Motor variables in the vocal tract and their reference activation levels, $u_r$, in the plosive-vowel sequences

| | OO | RO | GGa | GGp | HG | SG | MA |
|---|---|---|---|---|---|---|---|
| p/b | 1.0 | 1.0 | 0.0 | 0.0 | 0.5 | 0.0 | 0.0 |
| t/d | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| k/g | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| ɑ | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 |
| i | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| ɒ | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| u | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 |
| e | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 |

During on-line tracking, the E-FNN controller retrieves the muscular activations in the CV sequences and reproduces the desired articulatory trajectories. The reference MVs in Table 3.1 will be updated by the PID compensator on the phonetic segments. The off-line learning in this study is analogous to the babbling stage in human speech acquisition while the on-line tracking corresponds to the imitation stage (Bailly, 1997; Kröger et al., 2009). If act alone, the E-FNN can perform phoneme recognition tasks on the articulatory data, which is similar to the methods in (Richmond, 2009).

### 3.5.2 Off-line Training

The E-FNN learning algorithm operates at a sampling rate of 100 Hz. The MOCHA training data are divided randomly into five sets, four of which are used

for off-line training, the other one used for on-line tracking. The structure and parameter of the E-FNN are determined simultaneously on the four training sets of MV and TV data pairs. Using the learning algorithm, a total number of 23 fuzzy rules are created after training. Fig. 3.4 shows that the root mean square error (RMSE) converges after training on 250 samples. Unlike fix structured ANNs, some fuzzy rules in E-FNN share the same membership function, which improves system efficiency (Gao and Er, 2005).

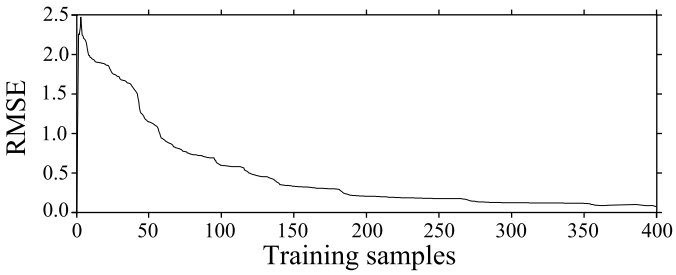



Figure 3.4: Average RMSE rate of the fuzzy neural controller on MV inversion using the MOCHA training data.

### 3.5.3 On-line Tracking

The trained E-FNN estimates the MVs given the desired articulatory trajectories. It couples with the PID compensator for on-line adaptive control of the MSD system. The controller manipulates the MSD to infer the muscular activation patterns and to reproduce the desired articulatory trajectories in the 2-D articulatory synthesizer. Mermelstein's 2-D vocal tract model with a tract length of 17.5 cm is divided into 89 tube sections, with a uniform length of $\Delta x = 1.966 \times 10^{-3} m$ and a thickness of $\Delta y = 1 \times 10^{-2} m$ (Boersma, 1998; Mermelstein, 1973). The tube wall property is measured in the relax cheeks of a human adult speaker (Birkholz and Jackel, 2004; Ishizaka et al., 1975), $M_0 = 21 kg/m^2, B_0 = 8000 kg/m^2 s, K_0 = 845000 kg/m^2 s^2$. The mass, damping,

and stiffness parameters of the MSD manipulator in (3.1) are calculated as:

$$M = M_0 \Delta x \Delta y = 4.129 \times 10^{-4} kg, \tag{3.21}$$

$$B = B_0 \Delta x \Delta y = 0.157 kg/s, \tag{3.22}$$

$$K = K_0 \Delta x \Delta y = 16.615 kg/s^2. \tag{3.23}$$

Initial conditions are set as $\ddot{z} = 0$, $\dot{z} = 0$, & $z = 0$, when reproducing the CV sequences. The gains of the PID compensator are set as $K_p = 25$, $K_i = 30$, & $K_d = 5$. The learning rate is $\eta = 0.005$.

#### 3.5.3.1 Articulatory Trajectories

Smoothness is a main property of human speech articulation. Thus the reproduced articulatory trajectories in the proposed controller are compared with the recorded EMA data of human speakers. Table 3.2 summarizes the RMSE of the controller during on-line tracking. The controller is able to manipulate the MSD and reproduce the desired position and velocity trajectories with high accuracy. Some TVs such as LI, TT, and TR demonstrate relatively higher error rates than others in Table 3.2. The observation suggests that the alveolar and the velar plosives possess a large amount of uncertainties in the CV sequences. In articulatory synthesis, the plosives are often independently generated using additional energy source at the constriction of the vocal tract during acoustic modeling (Birkholz et al., 2011). In practice, many researchers have suggested to reduce the degree of freedom or the error-prone TVs to increase the system efficiency (Birkholz, 2005; Ogata and Sonoda, 2003). For example, the jaw movement is often considered as a secondary feature that smooth the formant patterns during vowel production.

Treated as an inversion mapping module, the E-FNN is comparable to the neural-based trajectory mixture density network (TMDN) of Richmond (2009) and the Gaussian mixture model-based maximum likelihood estimation (MLE) in Toda et al. (2008). Table 3.2 shows that the E-FNN controller obtains similar RMSE results on the position parameter ($z$) of the selected articulatory parameters. Its overall error $1.608mm$ is slightly lower than MLE $1.413mm$ and TMDN $1.555mm$. However, E-FNN emphasizes on the joint accuracy of both the position and the velocity parameter, which is not available in previous studies. In this

study, the retrieved MVs (position and velocity parameters) provide an additional or alternative set of distinctive features to describe the variability of the surface acoustic-phonetic events, which are useful for the next-stage speech recognition. Similar to TMDN and MLE, the proposed neural model also uses a low pass filter to smooth the APFs to eliminate unrealistic or abrupt articulatory movements (cf. Section 5.4, Chapter 5).

Table 3.2: RMSE of the estimated articulatory trajectories in comparison with the EMA recordings.

| | E-FNN | | MLE | TMDN |
|---|---|---|---|---|
| | z (mm) | $\dot{z}$ (mm/s) | z (mm) | z (mm) |
| $UL_x$ | 0.67 | 0.52 | 0.76 | 0.91 |
| $UL_y$ | 1.20 | 0.81 | 1.05 | 1.06 |
| $LL_x$ | 0.75 | 0.66 | 1.12 | 1.12 |
| $LL_y$ | 1.04 | 0.98 | 1.90 | 2.22 |
| $LI_x$ | 2.07 | 1.36 | 0.69 | 0.81 |
| $LI_y$ | 1.45 | 1.13 | 1.02 | 1.03 |
| $TT_x$ | 1.65 | 1.02 | 2.07 | 2.10 |
| $TT_y$ | 2.53 | 1.64 | 2.24 | 1.94 |
| $TB_x$ | 1.97 | 1.56 | 1.94 | 1.98 |
| $TB_y$ | 1.90 | 1.58 | 1.82 | 1.78 |
| $TR_x$ | 2.04 | 1.74 | 1.86 | 1.83 |
| $TR_y$ | 2.03 | 1.55 | 1.90 | 1.88 |
| Average | 1.608 | 1.213 | 1.413 | 1.555 |

#### 3.5.3.2 Muscular Activations

For adaptive control, it is beneficial to extract the underlying articulatory commands or the MVs to explain the dynamics of the TVs. Fig. 3.5 shows the motor control signals of two MVs, OO and HG, in the proposed controller for the reproduction of [bɑ] sequence. Fig. 3.5 (a) and (d) plot the reference control signal $u_r$ (dashed line) and the inferred control signal $u_c$ (solid line), where

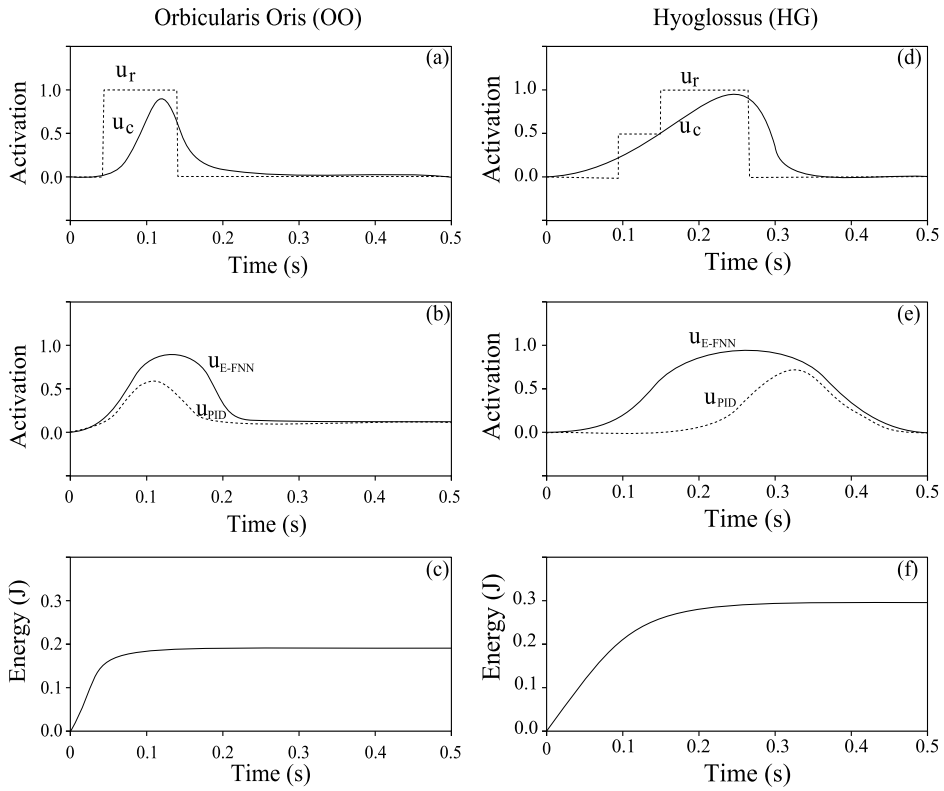



Figure 3.5: The characteristics of motor activation and energy consumption in the OO and HG during [bɑ] reproduction. Upper panels (a) and (d) plot the reference control signal $u_r$ (dashed line) and the inferred control signal $u_c$ (solid line) in the fuzzy neural controller; Middle panels (b) and (e) plot the control signal of the E-FNN, $u_{E-FNN}$, and that of the PID compensator, $u_{PID}$, for the MSD system; Lower panels (c) and (f) plot the control effort or the energy consumption in the MSD system.

$u_c = u_{E-FNN} - u_{PID}$ (Section 3.4, (3.14)). Fig. 3.5 (b) and (e) plot the control input of the E-FNN, $u_{E-FNN}$ and that of the PID compensator, $u_{PID}$. The proposed E-FNN controller demonstrates better performance than the linear compensator in terms of inversion accuracy. The motor activation data agrees with the measured EMG data of Baer et al. (1988), where the phones have distinctive targets in the control space. Three things are observed in the data.

1. the motor activation resembles the step response in the MSD. Ogata and

Sonada have previously used time-invariant linear systems excited by impulse trains to reproduce the velocity profiles of the speech articulators, which resembles the idea of motor inversion in this study (Ogata and Sonoda, 2003).

2. the delay between the MV onset and the TV onset (at the boundary of the gate-like reference control signal) is roughly 30-70 ms, which corresponds to the reaction time from muscle activation to articulatory motion in human speech production (Birkholz et al., 2011).

3. the controller is able to model the non-linearities of the articulators during CV production, which is evidenced by the smooth HG activation curve from [b] to [ɑ].

In the MSD based vocal track model, the controller moves the articulator back to the neutral position without oscillation, which mimics the human speech articulation (Perrier and Ostry, 1996; Saltzman and Munhall, 1989).

## 3.6 Discussion

The control model can be integrated with the anatomical and the acoustic models in a full articulatory synthesizer for TTS applications. However, the control model needs to specify the prosody information besides the phone sequences, such as the speaking rate in the input text. One difficulty is that the mapping from the articulatory trajectories to the acoustic sound is not strictly *one-to-one*, where there can be more than one vocal tract-cord configuration that produce the same acoustic sound in the synthesis system. The issue can be simplified by balancing the trade-off between the articulatory effort and the acoustic distinctiveness using an optimal control strategy (Kröger et al., 2009; Perrier et al., 2005). For speech motor control, the EMG measures are the combined results of efferent and afferent influences from the bio-mechanical properties of the muscular structures (Buchaillard et al., 2009; Perrier et al., 2005). The muscular forces can alter the position and velocity of the articulators. In the proposed controller, we are able

to examine the excitation pattern and calculate the input energy of the dynamic MSD system,

$$E = \int_{t_0}^{t_s} u(t)\dot{z}(t)dt, \tag{3.24}$$

where u(t) is the input force, $u_c$, $\dot{z}(t)$ is the velocity of the tract wall at time t, and $t_s$ is the sampling time. Fig. 3.5 (c) and (f) plot the energy consumption in joules (J) of two MVs, OO and HG, during [bɑ] production. Energy rises abruptly for OO in the lips when producing the labial plosive [b], but the overall measure is lower compared to HG when producing the low vowel [ɑ].

In the proposed controller, it is possible to calculate the overall control effort of the MVs in the articulatory synthesizer, where a *minimum energy* criterion can be embedded for optimal control (Kawato et al., 1990). The criterion is analogous to the speaker-oriented *minimum articulatory cost* in the functional phonology of speech production, where the speaker seeks to minimize the articulatory effort while maintaining the distinctiveness of the acoustic sounds during speech production (Boersma, 1998; Browman and Goldstein, 1992). However, the control energy in Fig. 3.5 (c) and (f) are only relative measures since the MVs are normalized to [0, 1]. For example, the activation of GGp (=1) exerts a force of 25.82 N during [i] production, while the activation of SG (=1) exerts a force of 6.9 N (Buchaillard et al., 2009). If used for optimal speech motor control, the MVs should have different prominence. However, Perrier et al. argued that the optimized control is not necessary for the smoothly varying articulatory movements. They showed that the bio-mechanical characteristics of the speech articulators alone can answer for such kinematic property (Perrier et al., 2005, 2003). In the present study, the MVs are extracted for the 2-D articulator with MSD based tube wall. Therefore, it has limited capability when evaluating the optimal control strategies. Nonetheless, the proposed controller is the first step toward an automatically controlled articulatory synthesizer. The inferred MVs can also provide an alternative set of motor features to describe the acoustic-phonetic events for improved speech recognition.

## 3.7 Summary

The shape, position, and movement of the articulators are the immediate targets of human speech production (Kelso et al., 1986; Saltzman and Munhall, 1989). Reproducing these smooth and natural trajectories is critical for high quality articulatory speech synthesis. This chapter presents an adaptive fuzzy neural controller, which tracks the measured articulatory trajectories in the form of TVs and infers the underlying muscular excitation patterns in the form of MVs. Major characteristics of the proposed adaptive fuzzy neural controller are as follows.

1. The E-FNN controller models the inverse dynamics between the motor commands and the tract variables in an off-line mode, where the structure and parameters of the neural topology are automatically and dynamically determined on the speech data.

2. The controller deals the uncertainties and the non-linearities in the MSD system using an adaptive control law.

3. Compared to the fixed structured ANNs, the self-adaptation and learning ability of the E-FNN controller is more adequate to model the dynamics of the articulators.

The proposed controller demonstrates good tracking performance on the CV sequences. It reproduces the smooth and bell-shaped articulatory trajectories as observed in the EMA data, and it retrieves the motor activations patterns in the vocal tract similar to the EMG data.

# Chapter 4

# Articulatory Phonetic Analysis of English Speech

## 4.1 Overview

The phonetic events exhibit correlated yet distinctive properties in the acoustic, the articulatory, and the auditory domain. The concept was first introduced in speech synthesis, and was later used for speech recognition (Yu and Oh, 2000). The hypothesis is that the articulatory feature space presents a much smaller variance than the acoustic feature space. This chapter describes a multi-dimensional pronunciation modeling method to extract these trajectory features for improved speech recognition, as illustrated in Fig. 4.1. It introduces the non-uniform segments to represent the pronunciation variations in English speech [1]. It also applies the heuristic articulatory-acoustic mapping to project the highly variate acoustic signals onto a set of pronunciation models conditioned on the conjoined principles of human speech production and perception.

For reliable articulatory-acoustic mapping, we need parallel recordings of acoustic and articulatory signals, discrete or continuous, with accurate phonetic annotations. Though there are various collections of direct physiological measurements available, e.g., the multichannel articulatory (MOCHA)-Texas Instrument

[1] The original manuscript was revised and re-submitted to IEEE Transactions on Audio, Speech, and Language Processing in June 2012. Earlier results of the research were presented and published in (Huang and Er, 2010, 2012b,c,d).

and Massachusetts Institute of Technology (TIMIT) corpus (Wrench, 1999), they are usually too limited for automatic speech recognition (ASR) studies due to the difficulty of data collection. Moreover, these corpora are generally speaker-dependent and target-specific, which render them too sparse. On the other hand, the analysis by synthesis approach has proved to be extremely useful when dealing with unfamiliar speech events in speech synthesis and speech recognition, such as out-of-vocabulary words and pronunciation variations. For instance, speech synthesis models have been used to search the acoustic correlates for the highly error-prone phoneme classes in ASR, such as stops (Blumstein et al., 1977), fricatives (Heinz and Stevens, 1961; Hughes and Halle, 1956), and nasals (Liberman, 1957; Malecot, 1973; Recasens, 1983). In fact, the synthesis method offers a more explicit and much simpler explanations of the correlation between the articulatory gestures and the acoustic realizations. To this end, a bio-mechanical speech synthesizer with closely monitored physiological properties is used to represent the average human adult speaker. Then the synthesizer is equipped with the English pronunciation prototypes and their diacritic variations extracted from an extensive hand labeled speech database. Finally the rule-based heuristic learning algorithm is applied to optimize the mapping between the pronunciation models and the articulatory configurations.

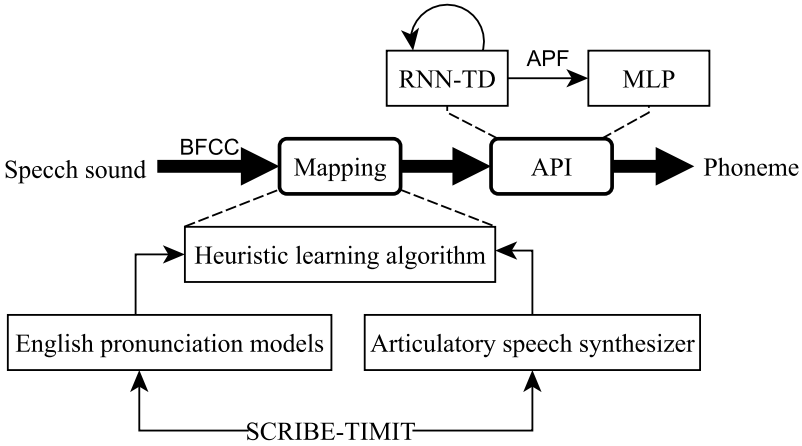



Figure 4.1: Schematic overview of the proposed multi-dimensional pronunciation modeling method for phoneme recognition.

The chapter is organized as follows. Section 4.2 introduces the bio-mechanical synthesizer, including the mathematical formulation of the human vocal system and the physiological components. Section 4.3 describes the proposed multi-dimensional pronunciation modeling method in detail. Section 4.4 simulates the articulatory-acoustic mapping using example consonant vowel (CV) patterns. Section 4.5 discusses the observations on the simulation results of the CV patterns, and elaborates on the bi-directionality of human speech. Section 4.6 concludes the chapter.

## 4.2 Articulatory Synthesizer

This section describes the dynamics in an articulatory synthesis system based on a mathematical formulation of the human vocal system from the glottis to the lips. The articulatory synthesis system has three physiologically derived components: the anatomical/geometric structure (4.2.1), the acoustic wave propagation (4.2.2), and the gestural control (4.2.3). The bio-mechanical system of Boersma (1998) is used here to study the motor control of the acoustic and the articulatory models. Boersma's articulatory synthesizer specifies the vocal apparatus from the lungs to the lips as air-filled tube sections which have elastic walls analogous to the damped mass-spring system, or the mass-spring damper (MSD). It simulates the articulatory and the acoustic equations simultaneously and systematically, which echoes the myeo-elastic and aero-dynamic theory of phonation (Bickford, 2006; Titze, 1980). The synthesizer is able to objectively and forcefully span the articulatory space by bringing the muscular apparatus from their equilibrium, maximum, and minimum positions toward the target region. The resulting acoustic outputs coupled with the articulatory gestures thus form an overall map of the many-to-one correlation between the the articulation configurations and the acoustic outputs.

Speech is synthesized with the 89 tube-sectioned bio-mechanical model, which consists of 28 muscular structures in 12 major groups, as summarized in Table 4.1, with reference to (Boersma, 1998; Mermelstein, 1973). The controlling parameters are named after the corresponding muscles for simplicity and clarity. The

muscular groups assimilate the human anatomy in physiological and functional properties of speech production (Boersma, 1998), as shown in Table 4.2.

Table 4.1: 28 controlling parameters in the bio-mechanical speech synthesizer.

| Major muscular groups | Controlling parameters |
|---|---|
| Subglottal system | *lungs* |
| Intrinsic laryngeal | *interarytenoid*, *cricothyroid*, *thyrovocalis*, *thyroarytenoid*, *posterior cricothyroid*, *lateral cricothyroid* |
| Extrinsic laryngeal | *stylohyoid*, *sternohyoid* |
| Lower pharynx | *stylopharyngeus* |
| Epiglottis | *sphincter* |
| Upper pharynx | *lower/middle/higher constrictor* |
| Palate | *levatorpalatini*,*tensorpalatini* |
| Tongue body and root | *verticalis*, *transversus*, *styloglossus*, *hyoglossus*, *genioglossus* |
| Tongue tip | *upper/lower tongue* |
| Jaw | *masseter*, *lateralptergoid* |
| Lip | *risorius*, *orbicularisoris* |
| Cheek | *buccinator* |

Below the larynx, the respiratory tract has 17 levels of branches between the trachea and the bronchioles. The number of branches almost doubles at each level, so there are $2^{17}$, or about $130,000$ of respiratory bronchioles. Each of the first order respiratory bronchioles supplies the primary lobule, which contains about two thousand alveoli, about 3.5 mm in diameter. At the deepest level, the respiratory bronchioli are connected to 300 million alveoli. These specifications of the sub-glottal systems yield realistic values for the pulmonic energy source. When the speaker inhales before articulation, the air pressure in the lungs raises to 954 Pa relative to the atmosphere pressure with expanded lungs and closed glottis. When the speaker releases the inspiratory muscles during articulation,

Table 4.2: Major muscular groups in the bio-mechanical synthesizer and their physiological properties.

| Major muscular groups | Physiological properties |
|---|---|
| Subglottal system<br>Supraglottal cavities | Generation of potential energy, i.e., air pressure<br>Transformation of kinetic energy, i.e., dynamic airflow |
| Larynx | Glottal phonation through opening and closing of the glottis |
| Velopharyngeal port | Regulation of air stream between the oral and nasal cavities |
| Tongue | Regulation of velaric air stream by changing shapes of the vocal tract |
| Jaw, lip, and cheek | Regulation of air pressure between oral cavity and the outside atmosphere |

the air pressure in the lungs eventually decreases and stabilizes at about 600 Pa. The volume change of the vital capacity by 10% is close to the pulmonic relaxation curve of human speakers (Boersma, 1998; Hixon, 1987). In the larynx, i.e., the voice box, the extrinsic laryngeal muscles have one attachment point outside the larynx, and the intrinsic laryngeal muscles have both attachments within the larynx, the coordination of which can realize a pitch range from $50Hz$ up to $700Hz$ which is close to the human capacity (Nunn, 1993).

Another closely modeled critical articulator is the tongue centered in the vocal tract above the larynx. For human speech motor control, the tongue covers many, if not most, of the functional aspects of phonation. Majority of the tongue muscles are paired symmetrically on each side of the mid-sagittal plane. This structure makes the tongue extremely flexible serving its primary purpose of food mastication and the secondary function of sound articulation. In the synthesizer, the tongue has four intrinsic muscles which change the shape of the tongue body, i.e., superior longitudinal (SL), inferior longitudinal (IL), verticalis (VS), and transversus (TS), and four extrinsic muscles which change the position the tongue body, i.e., genioglossus (GG), hyoglossus (HG), styloglossus (SG), and

palatoglossus (PG), where the coordinations of GG, HG, and SG are critical for vowel production. For example, the GG is the largest muscle in the tongue, and controls the front-back position of the tongue. And a compression force on the GG will induce a forward displacement of the tongue body and a slight elevation of the upper part of the tongue, and vice versa. The resulting tongue shapes are mainly observed for high front vowels such as [i]. In practice, the high front vowels also require further compression of the sides on the tongue through the HG muscle, which consists of two rectangular-like parts on each side of the tongue body. In functional phonology, the height of English vowels are mainly controlled by the GG and HG configurations, and could be further refined or adjusted to enhance audibility by the secondary movements such as the jaw height, which is mainly controlled by the *masseter*. The group of SG fibers are located inside the tongue body laterally. Positive forces produced by the SG bunch the tongue body backward, and elevate the tongue root toward the velar region, which result in a backward displacement of the tongue body and a lowered tongue tip. This resembles the production of velar sounds such as the high back vowel [u] and the plosive [k]. The former requires lip rounding which is controlled by *risorius* and *orbicularisoris*, and the latter indicates constriction between the SG and the velar (Badin et al., 2002; Birkholz et al., 2011).

The meshed supra-glottal, pharyngeal and oral cavities of the resulting biomechanical model in a neutral position has the articulatory configuration of an average human adult speakers. For example, the physiological factor $f = 1.2$ in the synthesizer produces an average vocal tract length of 16.9 cm which is close to an adult male, and $f = 1.0$ results in a vocal tract length (VTL) of 14.1 cm for an adult female (Bickford, 2006). The average length of the tongue from the oropharynx to the tip is approximately 10 cm, similar to human. And the lips are able to protrude/spread to prolong/shorten the VTL by $1 - 2$ cm as human speakers do (Bailly, 1997; Bailly et al., 1997). In this study, the muscular parameters (e.g., activation levels) that control the MSD based tube walls are the motor variables (MVs). The articulatory parameters (e.g., position and velocity trajectories) of the corresponding apparatus (e.g., the tongue body) are the tract variables (TVs) (cf. the control scheme in Chapter 3).

### 4.2.1 Soft-body Dynamics: Anatomical Model

Human vocal system resembles a volumetric deformable duct, the walls of which include a mixture of organic materials such as muscles, fat, and bones. The articulatory synthesizer of Boersma (1998) expands Mermelstein's 2-D vocal tract structure, which is based on the X-ray tracings of a human speaker (Mermelstein, 1973). The vocal system is modeled as concatenated tube sections filled with air, the walls of which are modeled as the MSD. The method allows the wall to yield to air pressure changes during the acoustic simulation (Boersma, 1998). Each tube section has an uniform length $\Delta x$. The meshed supra-glottal, pharyngeal and oral cavities of the resulting anatomical model in a neutral position has a VTL of 16.9 cm. The average length of the tongue from the oropharynx to the tip is approximately 10 cm. And the lips are able to protrude/spread to prolong/shorten the VTL by $1-2$ cm (Boersma, 1998). The 8 MVs include superior longitudinal (SL), anterior genioglossus (GGa), posterior genioglossus (GGp), hyoglossus (HG), styloglossus (SG), masseter (MA) , risorius (RO) , and orbicularisoris (OO). The muscular structures are independently controllable (Bouabana and Maeda, 1998). The 12 TVs include the x-y coordinates of the tongue root (TR), the tongue body (TB), the tongue tip (TT), the lower lip (LL), the upper lip (UL), and the lower incisor (LI) (cf. Section 3.3 in Chapter 3). Each articulatory parameter represents one degree of freedom. Though the degree of freedom should be as many as possible to approximated the human system, it should be as few as possible to avoid exponential increase of the tract shapes in the articulatory synthesizer (Birkholz et al., 2007; Perrier et al., 2003). The selected TVs have already shown promising results for consonant-vowel synthesis (Buchaillard et al., 2009; Perrier et al., 2005) and for phoneme recognition (King et al., 2007; Mitra et al., 2011).

### 4.2.2 Fluid Dynamics: Acoustic Model

Acoustic wave propagation inside the vocal tube/duct follows two physic laws: the equation of motion (i.e., the conservation of momentum) and the equation of

continuity (i.e., the conservation of mass),

$$A(x)\frac{\partial P(x,t)}{\partial x} + \rho\frac{\partial U(x,t)}{\partial t} = 0, \tag{4.1}$$

$$\frac{\partial U(x,t)}{\partial x} + \frac{A(x)}{c^2}\frac{\partial P(x,t)}{\partial t} = 0, \tag{4.2}$$

where A(x) is the cross-sectional area of the tube section at position x, P(x,t) is the pressure at position x at time t, U(x,t) is the volume velocity past position x at time t, $\rho$ is the air density, and c is the speech of sound in air. Since the area function is constant, $A_m(x) = A_m$, in a tube section $m$, Webster's horn equations can be derived as:

$$\frac{\partial^2 P_m(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 P_m(x,t)}{\partial t^2}, \tag{4.3}$$

$$\frac{\partial^2 U_m(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 U_m(x,t)}{\partial t^2}, \tag{4.4}$$

for the pressure and the volume velocity in the $m^{th}$ tube section respectively.

To solve the partial differential equations, there are mainly three types of methods or acoustic models for the sound simulation: the Kelly-Lochbaum (KL) model (Kelly and Lochbaum, 1962), the transmission line model (TLM) (Maeda, 1982), and the hybrid time-frequency domain method (Sondhi and Schroeter, 1987). These articulatory synthesizers separate the vocal tract into a source and a filter part in the acoustic model. The TLM and the KL model are based on the analogy between the acoustic tubes and the electrical circuits. The TLM uses varying tube length, which is not accounted for in the KL model. However, neither of them models the yielding tube walls due to the air pressure. For example, van den Doel and Ascher (2008) placed a separate voice source (in the two-mass glottal model of Flanagan (1972)) and a wall vibration model to address the above weaknesses of the KL model.

The source-filter separation models are able to produce smooth and natural vowel sounds much more effectively than the consonantal effects. Moreover, they all have trouble simulating myoelastic-aerodynamic interactions between the vibrating vocal folds and the resonating vocal tract (Boersma, 1998; Mermelstein, 1973; Titze, 1980). One salient property in the two-mass glottal excitation model

of Flanagan (1972) is the interaction between the voice source and the vocal tract. The two-mass glottal excitation model specifies three controlling parameters: the mass, the spring, and the subglottal pressure. In the articulatory synthesizer of Boersma (1998), the two-mass model was used to construct the walls of the entire vocal duct. The effects of varying tube length, yielding tube walls, and turbulent noises are implicitly included during the numerical simulations of the partial differential equations.

This study focuses on modeling the vocal tract above the glottis. In previous chapter, the non-linearity and the dynamics of the vocal system was addressed using the adaptive neural controller. The behavior of glottal excitation in voiced/unvoiced speech production is also nonlinear. The glottal model is driven by the lung pressure and the glottal width induced by the tension in the vocal cords. It requires logical decisions, which are the most costly when implementing the signal processor (Cook, 1990; Lippmann, 1987). Moreover, the airflow is not always laminar in actual articulation due to the viscosity of the fluid. Turbulent noises are generated at the narrowest constriction in the vocal tract, which corresponds to the place of articulation for the consonants, e.g., fricative and plosives. In this study, Reynolds number is used to indicate the viscous force within the fluid at the constriction,

$$Re = \frac{2U}{v\sqrt{A\pi}}, \tag{4.5}$$

where A is the area of the constriction aperture, U is the volumetric flow through the aperture, and $v$ is the ratio of the dynamic viscosity against the density, $v \approx 0.15 cm^2/s$ for dry air. Turbulence occurs when the Reynolds number exceeds a critical value, $Re_c$. The burst frequency (BF) in the spectrum of the turbulent noise is,

$$f = \frac{SU\sqrt{\pi}}{2\sqrt{A^3}}, \tag{4.6}$$

where S is the Stouhal number (Cook, 1990).

The resulting acoustic sound pressure in $Pa$ $(N/Mm^2)$ at a distance $d$ from the lips is (Boersma, 1998),

$$P(t,d) = \frac{4\pi\rho_0}{d}\left[c \cdot \frac{\partial\left(A_{89}(t)\Delta y_{89}(t)\right)}{\partial t^2} + \sum_{m}^{m=M} 1000\Delta x_m \Delta z_m \frac{\Delta y_m}{\partial t}\right], \tag{4.7}$$

where $M = 89$ is the total number of tube sections in the meshed model, $\rho_0 = 1.14$ $kg/m^3$ is the reference air density inside the vocal organs, $A_{89}(t)$ is the lip area that modifies degree of lips rounding, $\Delta y_{89}(t)$ measures the degree of of lips protruding, and $c$ accounts for the air leak through the nose. The acoustic sound of the synthesizer is the linear superposition of the lip radiation, the nasal radiation and the turbulent noise, which are observed in plosives, nasals, and fricatives. The acoustic sounds are sampled at $22,050$ Hz, whereas the articulatory configurations are sampled at 100 Hz. The aerodynamic and myoelastic transformations are computed 100 times every second.

### 4.2.3 Articulatory Targets: Control Model

The widely used control input is the articulatory gestures that are derived in the task-dynamic approach (Browman and Goldstein, 1992; Saltzman and Munhall, 1989). Each phone or phonetic sequence is specified by a set of gestures, or gesture scores (Kröger et al., 1995; Saltzman and Munhall, 1989). Birkholz et al. (2007) used the control model to transform the gestural scores into a sequence of vocal tract and vocal fold parameters, which are similar to the TVs in Chapter 3. TVs can be directly transformed into the discrete tube geometry (2-D or 3-D) of the synthesis system (Birkholz et al., 2007; Kröger et al., 1995). TVs have also been used as articulatory representations of the phonetic events for speech recognition system (Mitra et al., 2011). Another type of control input is the muscular activations, which are based on the equilibrium point hypothesis (EPH) hypothesis or the $\lambda$-model (Feldman, 1986). The muscles are the force generators in this type of control input. Perrier et al. (2005) used the $\lambda$-model to generate all possible motor commands in a 2-D bio-mechanical synthesis model. They applied the overall map as an articulatory codebook, and used the radial basis function-based neural networks (RBF-NNs) based inversion model to map the feature vectors between the motor commands, the tongue shapes, and the acoustic sounds. Buchaillard et al. (2009) used the $\lambda$-model in a 3-D tongue structure to examine the impact of muscular control on the tongue shape and the generated French vowel sounds.

In the task-dynamic approach of Saltzman and Munhall (1989), the immediate targets: the shapes and positions of the articulators are fed to the articulatory synthesizer. One target per phone is specified. Target interpolation and approximation methods are used to generate the articulatory movements, e.g., the position and velocity profiles, in the phonetic sequences (King et al., 2007; Perrier and Ostry, 1996; Perrier et al., 2003). In the EPH approach of Feldman (1986), the muscular activities: the activation force in the articulators are fed to the articulatory synthesizer. The dynamics of the articulators are usually modeled by linear $2^{nd}$ order system in analogy to the MSD system (Buchaillard et al., 2009; Perrier and Ostry, 1996). The shifts of the equilibrium positions of the speech articulators induce the articulatory movements (Feldman, 1986). In this sense, the EPH approach resembles the target approximation method (Birkholz et al., 2007). In TTS applications, the task-dynamic approach allows easy manipulation of the speech targets. However, Kröger et al. (1995) pointed out that the articulatory targets are not well fitted when compared to the smoothly varying trajectories of human speakers. They introduced a time-variant force function to approximate the transitional changes of the articulatory trajectories at phonetic boundaries (Kröger et al., 1995). In contrast, others used a higher order dynamical system to reproduce the articulatory trajectories (Birkholz et al., 2007; Ogata and Sonoda, 2003).

In Boersma's bio-mechanical synthesizer, the input to the control model are the MVs, which control the length and tension of the muscles (Boersma, 1998). The muscles form the walls of the vocal ducts from the lungs to the lips. The wall displacement caused by the muscular forces follow the $2^{nd}$ order equation in the MSD system . In the dynamic control model, the MSD is critically damped, which avoids target overshoot in the vocal tract movements (Kelso et al., 1986). The tube walls of each articulator are described by width $\Delta x$, length $\Delta y$, and depth $\Delta z$. These muscular groups assimilate the human anatomy both in physiological properties and in functionality of speech production. The articulatory model is configured to resemble an average human speaker (physiological factor $f = 1.1$), whose characteristics are listed in Table 4.3. The damping factor is 0.945 for an open wall in the tongue apparatus, and it assumes a critical value of 1 for closed walls to model the extra stiffness, which approximates a modal voice

with a normal rate of speech to agree with the Spoken Corpus Recordings in British English (SCRIBE) recordings (Perrier and Ostry, 1996). Timing in speech production requires a higher level of gestural activation model which are capable of intrinsic coordination of phonetic sequences (Boersma, 1998; Saltzman and Munhall, 1989). By specifying the start and the end time of muscular activities in the control model, the timing of the phones and the phonetic boundaries are explicitly defined in the pilot study of CV sequences. In the natural speech database, the timing information is extracted from the phonetic label file.

Table 4.3: Parameter settings of the articulatory based speech synthesizer.

| | Value | Unit |
|---|---|---|
| Sizing factor: $f$ | 1.1 | 1 |
| Tissue wall thickness: $\Delta y$ | 1.0 | cm |
| Tissue wall mass density:$\rho$ | $1.0 \times 10^1$ | $kg/m^2$ |
| Linear spring constant:$k_1$ | $1.0 \times 10^6$ | $N/m^3$ |
| Cubic spring constant:$k_3$ | 0 | $N/m^3$ |
| Linear tissue stiffness:$s_1$ | $5.0 \times 10^6$ | $N/m^3$ |
| Cubic tissue stiffness:$s_3$ | $2.5 \times 10^{13}$ | $N/m^3$ |

## 4.3 English Pronunciation Modeling

### 4.3.1 Pronunciation Models

A set of 255 pronunciation models are used to annotate the articulatory-acoustic data from a parental speech corpus, i.e., the SCRIBE-TIMIT dataset (Hieronymus et al., 1990; Millar et al., 1994). In this stage, the bio-mechanical articulatory synthesizer is coupled with the parental corpus, which consists of 100 phonetically rich TIMIT SX sentences (Garofolo et al., 1993) repeated by 5 native English speakers, 4 male (MAC, MAE, MAF, MAM) and 1 female (FAA), with the received pronunciation (RP) accent. There are two reasons that the SCRIBE-TIMIT corpus is chosen. On the one hand, SCRIBE-TIMIT gives 255 detailed phonetic annotations by trained professional linguistics, which clearly distinguish

the true base-forms of English phonemes from their allophonic variation using the extended Esprit Speech Assessment Methodology Phonetic Alphabet (X-SAMPA) labeling scheme encoded in ASCII. In contrast, while the conventional 61 TIMIT annotations encoded in Arpabet may be sufficient for experienced human speakers and listeners, they often indulge problematic assumptions, e.g., the independence of phonetic features and the quasi-stationarity of speech sounds. On the other hand, the SCRIBE-TIMIT corpus preserves meaningful knowledge sources such as the minimum audible changes of phonetic qualities, and it balances the trade-off between the perceptual invariance of human listeners and the articulatory effort of human speakers on the same set of sentences. These details render the selected corpus more descriptive the TIMIT database and more compact than the physiological database such as the electromagnetic articulograph (EMA) recordings for multi-dimensional phonetic representations.

CV patterns have been widely used as a testbed for speech recognition and synthesis experiments. They are the basic building blocks which are more descriptive than isolated monophones, and at the same time more compact than the composed words and sentences. Furthermore, they preserve the dynamics of natural speech including the coarticulations effects in the consonantal onset, the transitional region, and the vowel nucleus in great detail. This simulation study focuses on two sets of natural speech sounds. The first set consists of the six plosives: [b, p, d, t, g, k], representing the three pairs of voicing contrast with constriction at the lips: [b/p], the alveolar: [d/t], and the velar: [g/k], along the tongue continuum. The second set consists of the four primary cardinal vowels: [ɑ, i, ɒ, u], representing the four coordinate positions in the vowel chart, and schwa: [ə], representing the neutral vowel sound.

Table 4.4 shows the values of six MVs: OO, RO, GGa, GGp, HG, SG, and MA, at the target articulatory positions for the base-forms of 6 plosives, i.e., [b, p, d, t, k, g], and 5 primary cardinal vowels i.e., [ɑ, i, ə, ɒ, u]. The maximal and minimal moving ranges of the muscle width are normalized to be in the $[-1, +1]$ interval. The range is equivalent to that of the normalized MVs $[0, 1]$ in Table 3.1 (cf. Chapter 3). The negative interval is used here to give a clear representation in the articulatory domain. A value outside this range indicates extra compressive or de-compressive myoelastic tension of the tissues but does not introduce further

vocal tract deformation. For the highly elastic tongue muscles, the boundary values would result in constriction at certain parts of the vocal tract walls. In previous works, the extreme articulatory conditions are also referred to as the virtual target, since the tongue does not actually reach these extreme positions outside the vocal tract (Birkholz et al., 2011). The controlling parameters are set according to the phonological definitions of the phonemes and the X-ray image of human objects during speech productions (Jones, 1972). The standard phonetic base-forms using international phonetic alphabet (IPA) annotations, the broad phoneme form using the Arpabet, and the narrow phonetic form using X-SAMPA annotations are listed side by side. For simplicity, IPA symbols will be shown in square brackets, Arpabet in forward slashes, and X-SAMPA in double slashes when they appear. The full listing of the X-SAMPA symbols is available from (Wells, 1997).

For the three types of plosives, i.e., bilabial [b, p], alveolar [d, t], and velar [g, k], there are two stages of articulation, closure and release, where the distinction between voiced plosives [b, d, g] and their unvoiced counterparts [p, t, k] depends on the intervocalic period from the lip closure to the vocal fold vibration, i.e., VOT, which is the voice onset time (VOT) . In general the VOT threshold of voiced-unvoiced separation has the highest boundary value for velar sounds, and the lowest for bilabial sounds. In this study, the cut-off VOT thresholds for the plosives at syllable-initial positions are assigned with a 10 ms separation, e.g., the labial [b/p] at 30.16 ms, the alveolar [d/t] at 40.13 ms, and the velar [g/k] at 50.02 ms, according to (Stouten and Hugo, 2009; Zue, 2004). However, when the stops are not in syllable-initial position, the voicing cue is subject to the vocal fold vibration of the neighboring phonemes (Martin and Jurasfsky, 2008). For the example used here, the tongue muscles are prepared for the primary vowel [ɑ] during plosive release to demonstrate the effect of anticipatory coarticulations in CV production as seen in natural human speech.

Fig. 4.2 shows the articulatory configurations of the bio-mechanical synthesizer for the three pairs of plosives, i.e., [b/p], [d/t], and [g/k] in CV production. Fig. 4.3 shows the shape, position, and degree of constrictions of the vocal tract for the five primary vowels: [ɑ, i, ə, ɒ, u], according to the settings in Table 4.4. The jaw opening is in some way a secondary feature for vowels, since one is still

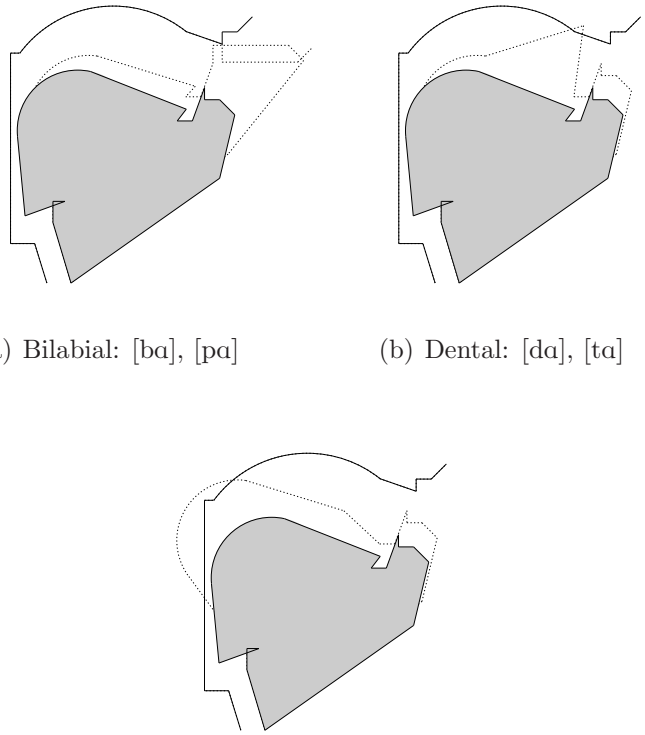
Table 4.4: Articulatory configurations of 7 muscles for 6 plosives and 5 primary vowels during CV production.

| | Base | Broad | Narrow | OO | RO | GGa | GGp | HG | SG | MA |
|---|---|---|---|---|---|---|---|---|---|---|
| Closure | b̥/p̥ | bcl/pcl | bv+bc/pv+pc | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | d̥/t̥ | dcl/tcl | dv+dc/tv+tc | 0.0 | 0.0 | 0.6 | 0.0 | 0.0 | 0.0 | 0.0 |
| | g̊/k̥ | gcl/kcl | gv+gc/kv+kc | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| Release | b̆/p̆ | b/p | ba/pa | 1.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 | 0.0 |
| | d̆/t̆ | d/t | da/ta | 0.0 | 0.0, | 0.6 | 0.0 | 0.5 | 0.0 | 0.0 |
| | ğ/k̆ | g/k | ka/ga | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 1.0 | 0.0 |
| Vowel | ɑ | aa | A | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 | -0.5 |
| | i | ih | I | 0.0 | 0.0 | 0.0 | 0.7 | -1.0 | -0.7 | 0.0 |
| | ə | ax | @ | 0.0 | 0.0 | 0.0 | -0.5 | 0.9 | -0.7 | 0.0 |
| | ɒ | ao | Q | 0.5 | -0.5 | 0.0 | -0.9 | 0.8 | -0.4 | 0.0 |
| | u | uw | U | 0.5 | -0.5 | 0.0 | 0.2 | -0.8 | -0.4 | 0.0 |

able to talk while holding the jaw but not so much while holding the tongue. Nevertheless, some linguists continue to use the degree of jaw opening to mark the vowel chart, where high vowels have closed jaws, and low vowels have open jaws. In this study, the MA is treated as a complementary feature, which actually increases the second formant value to smooth the acoustic trajectories, e.g., [ɑ]. Additionally the *risoris* is set to -0.5 to achieve lips rounding for [ɒ] and [u], the effect of which is not fully visible in the side view of the supra-glottal system. The velopharyngeal port, i.e., the velum, is closed for all these non-nasal sounds, where the *levatorpalatini* has default value 1.

### 4.3.2 Heuristic Learning Algorithm

The proposed heuristic learning algorithm is analogous to the human experience of speech acquisition through a language teacher. In other words, the bio-mechanical synthesizer is supervised by the native English speakers alongside the linguists in the SCRIBE corpus. The algorithm minimizes the within-class scatter distance, and maximize the across-class scatter distance for improved phonetic clustering. The maximal and minimal moving ranges of the controlling parameters, MVs,



(a) Bilabial: [bɑ], [pɑ]

(b) Dental: [dɑ], [tɑ]

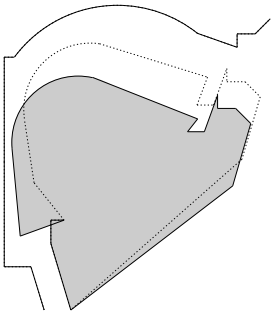
(c) Velar: [gɑ], [kɑ]

Figure 4.2: Shape and constriction of the vocal tract for three types of plosives: bilabial, dental, and velar. Dotted line: closure, solid line: release.

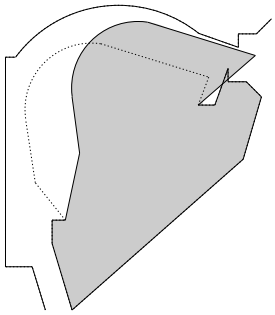
in the bio-mechanical speech synthesizer are normalized to $[-1, +1]$ interval. A value outside this range indicates extra compressive or de-compressive myoelastic tension of the tissues but does not introduce further vocal tract deformation. For the highly elastic tongue muscles, the boundary values would result in constriction at certain parts of the vocal tract walls. In previous works, the extreme articulatory conditions are also referred to as the virtual target, since the tongue does not actually reach these extreme positions outside the vocal tract (Birkholz et al., 2011).

The heuristic learning algorithm is defined as follows.

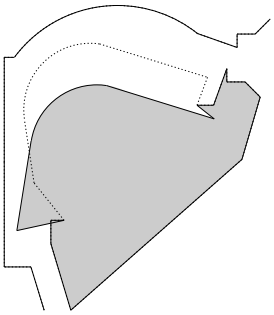
1. Let $i$ be the number of pronunciation models, where $i = 1, 2, \cdots, 255$;



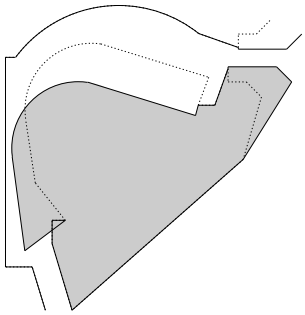
(a) Front-low-rounded: [ɑ]



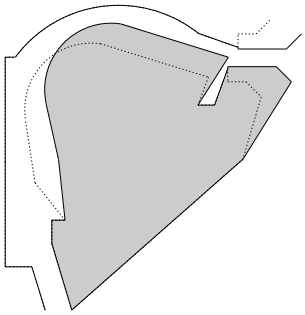
(b) Front-high-unrounded: [i]



(c) Schwa: [ə]



(d) Back-low-rounded: [ɒ]



(e) Back-high-rounded: [u]

Figure 4.3: Tongue configurations for the five primary vowels according to the articulatory target in Table 4.4. Dotted line: neutral position, solid line: target position.

2. Compute the mean $\mu_i(x)$ and the standard devision $\sigma_i(x)$ of the target auditory cepstral features in the phonetic segment;

3. Apply *Rule 1*, and compute the relative entropy $H(\mu, \sigma)$ of the synthesized and the target auditory cepstral features;

4. Apply *Rule 2*, and compute the production cost $C_{RAT}$ of the articulatory movements;

5. Set $y^i(t) = y^i(t) \pm d_y$, with $d_y = 0.1$ and $y \in [-1, +1]$;

6. There are two scenarios:

   (a) if $H(\mu, \sigma) < H(\mu, \sigma)|_{y \pm d_y}$ or $C_{RAT} > C_{RAT}|_{y \pm d_y}$, then repeat step 3 and 4;

   (b) if $H(\mu, \sigma) \geq H(\mu, \sigma)|_{y \pm d_y}$ and $C_{RAT} \leq C_{RAT}|_{y \pm d_y}$, then repeat step 3, 4, and 5 for $x^i$ and $z^i$;

   (c) else set $i = i + 1$, i.e., go to the next phonetic label and iterate for all phonetic segments.

The two heuristic rules are as follows.

1. *Rule 1:* Listener oriented optimization of acoustic qualities using the relative entropy measure: $H(\mu, \sigma)$. Since not all of the synthesized acoustic sound result in audible or meaningful outputs, the auditory features are extracted to evaluate the synthesized pronunciations. The auditory model integrated the 23 channel bandpass Gamma-tone filters spaced uniformly on the bark scale to extract the auditory feature vectors, bark-frequency cepstral coefficients (BFCCs). The frame rate 10 ms is chosen to match the sampling rate of the synthesized articulatory data. For feature decorrelation and dimensionality reduction, principle component analysis (PCA) is applied to the spectral sub-band energy vectors, which generates 12 static cepstral coefficients. The coefficients are augmented by their first and the second order derivatives. The derivatives (or delta) coefficients are calculated using two frames of past context and two frames of future context.

Log energy is attached resulting in 39 dimensional BFCCs. This procedure similar to the discrete cosine transform in the mel-frequency cepstral coefficients (MFCCs) of the baseline (Toda et al., 2008), For unification, the static PCA coefficients are augmented with delta and acceleration coefficients which result in the 39 dimensional feature vectors. The feature evaluation criterion is the relative entropy measure, $H(\mu, \sigma)$,

$$H(\mu, \sigma) = \sqrt{\sum_{i=1}^{39} (H_i^s - H_i^t)^2}, \tag{4.8}$$

where $H_i^s$ and $H_i^t$ are the entropy values of the $i_{th}$ coefficient of the synthesized and the target auditory cepstra, and

$$H_i = \sum_{x}^{no.of frames} \mu_i(x) log \frac{\mu_i(x)}{\sigma_i(x)}, \tag{4.9}$$

where $\mu_i(x)$ and $\sigma_i(x)$ are the mean and the standard deviation of the cepstral coefficients computed from all the time frames in the current phonetic segment.

2. *Rule 2:* Speaker oriented minimization of articulatory cost, $C_{RAT}$. In the synthesis model, the equilibrium position and the regional articulatory target (RAT) are defined to monitor the dynamic articulatory transactions, which have previously been used to improve the performance of speech synthesizers (Birkholz et al., 2011). During regional target approximation, the phonetic events are projected onto a spread region instead of singular points in the articulatory domain. The production cost is calculated as the articulatory effort required or the energy consumed for the muscle structure to move from the current place to the nearest point in the target region, or to return to the equilibrium position, defined as,

$$C_{RAT} = \frac{\partial W}{\partial E} \qquad = \frac{2a}{b} \times \frac{\partial (exp(b(I_1 - 3)) - p(I_3 - 1)^2)}{\partial (FF^T - I)}, \tag{4.10}$$

where $W$ is the stored strain energy, $E$ is the Lagrangian strain tensor, $I_1$ and $I_3$ are the first and the third invariant of the deformation tensor $E$,

where $I_1 = 3 + 2Tr(E), I_3 = det(2E + I)$, and $a$, $b$ & $p$ are tuning variables. The internal force $F$ in the muscle is defined as:

$$F = M\frac{\partial^2 \vec{V}}{\partial t^2} + B\frac{\partial \vec{V}}{\partial t} + K\vec{V}, \tag{4.11}$$

where $M$ is the mass matrix, $B$ is the damping matrix, $K$ is the elasticity matrix of the bio-mechanical synthesizer, and $\vec{V} = [x(t), y(t), z(t)]$ is the displacement vector. The minimization of the production cost keeps the variation of the muscular activities as low as possible near the target region, and it has a tendency of bringing the articulators back to equilibrium position, which is in accordance with human speech production (Perrier and Ostry, 1996).

The heuristic learning specification allows smooth articulatory assimilation, as well as easy explanations of many consonantal pronunciation variations. It iteratively span or shrink the articulatory target region to generate the distributed RATs for the pronunciation models in the resulting articulatory-auditory database. For each of the 255 phonemes in the validation set, the initial articulatory gesture is updated with a 5 ms frame rate to generate the most probable configurations that result in authentic acoustic qualities. In this manner, the two criterion iteratively spans or shrinks the articulatory target to generate a distributed RAT for the pronunciation models in the combined articulatory-auditory space.

## 4.4 Simulation on CV Patterns

### 4.4.1 Vowel Correlates

In the acoustic dimension, vowels are generally much more stationary than consonants. The first two formants in the fast Fourier transform (FFT) power spectrum, F1 and F2, to illustrate the multi-dimensional phonetic attributes of the CV patterns. In the articulatory dimension, the muscular activations of GG, HG, and OO are used. Fig. 4.4(a) shows the acoustic formant distribution of the vowel sounds using the standardized logarithm of the first two formants. Fig.

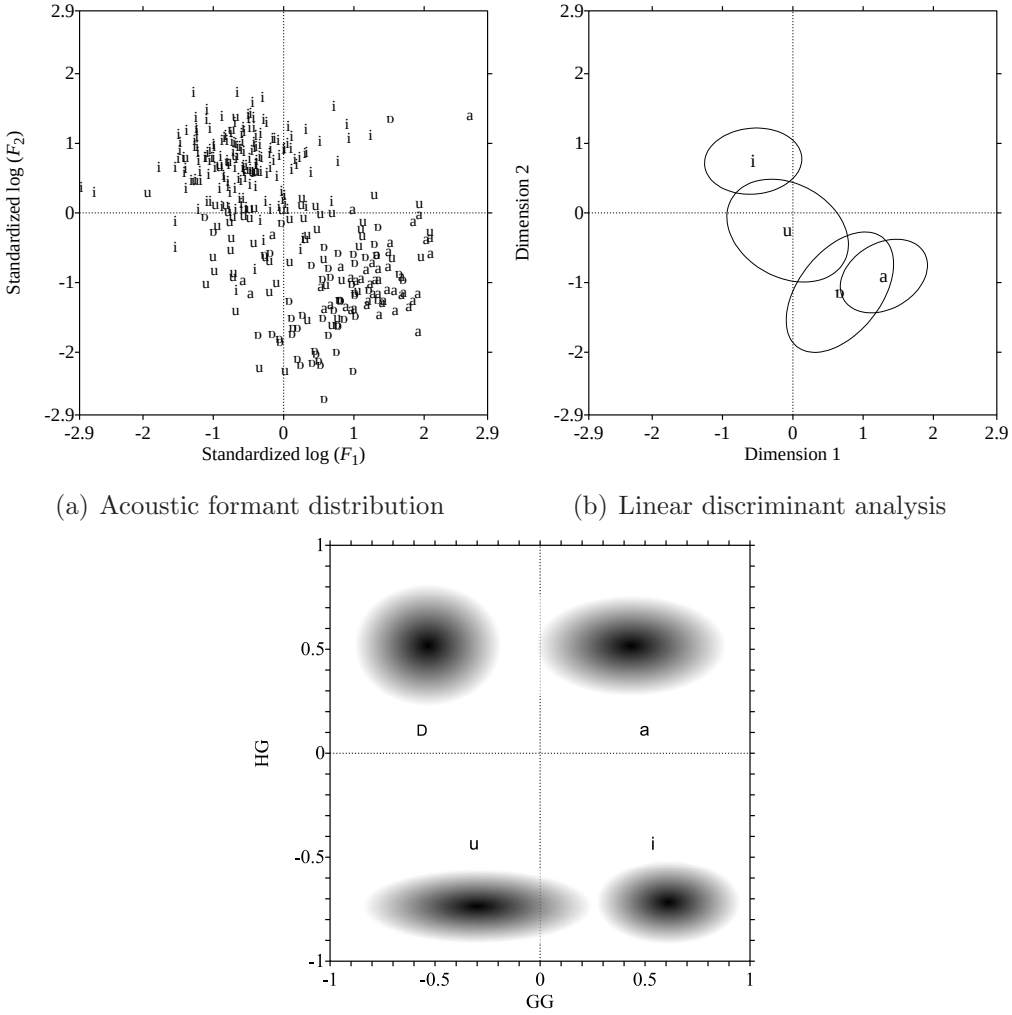
4.4(b) shows the auditory trajectory of the acoustic formants using linear discriminant analysis. It shows a classification ratio of 75.6% for the vowels. The reduced vowel, i.e., schwa: [ə], is excluded in Fig. 4.4(b). Fig. 4.4(c) illustrates the articulatory trajectory, i.e., two MVs: GG and HG. The RATs are shown as darkened gradient with one standard devision.

Two things could be observed. First, the mappings between the articulatory and the acoustic space on the phonetic patterns are not uniformly distributed, i.e., the *many-to-one* problem, which has been a major issue of the *analysis by synthesis* approach (Bickford, 2006). Secondly, the tongue proves to be a powerful and critical factor during continuous speech production. The muscle activities of each articulator on the tongue continuum can result in significant and audible change in the auditory space. The vowel height is inversely correlated to F1 value, e.g., the higher the F1 value, the lower or more open the vowel, thus the more intense the force in the HG muscle, and vice versa.

### 4.4.2 Consonant Correlates

For the consonantal onsets in the CV patterns, Fig. 4.5 shows the acoustic and articulatory trajectories of the three pairs of plosives, i.e., bilabial: [b/p], alveolar: [d/t], and velar: [g/k]. In particular, Fig. 4.5(a) shows the formant map for [b] and [p] in syllable initial pre-stressed positions in the SCRIBE-TIMIT sentences. Fig. 4.5(b) plots the discriminant analysis result of the [b/p] contrast using the formant measure, where a discriminant ratio of 56.3% is obtained. The ratio is much lower than that of the vowels as expected for the highly dynamic plosives. Fig. 4.5(c) illustrates the RAT distribution of the MVs: OO and GG for the plosive in the articulatory domain.

Compared to vowels, the plosives have much more compact definitions in the articulatory space, where VOT and other variant acoustic properties are simply caused by the precise timing of vocal fold vibration during articulation. In fact, VOT is often considered an eminent feature of plosives in speech production and perception. Fig. 4.6 shows the VOT distribution of the plosives in the recognition output of the SCRIBE-TIMIT sentences. It is observed that the perception of stops is rather categorical, where the boundary conditions of VOT value seems



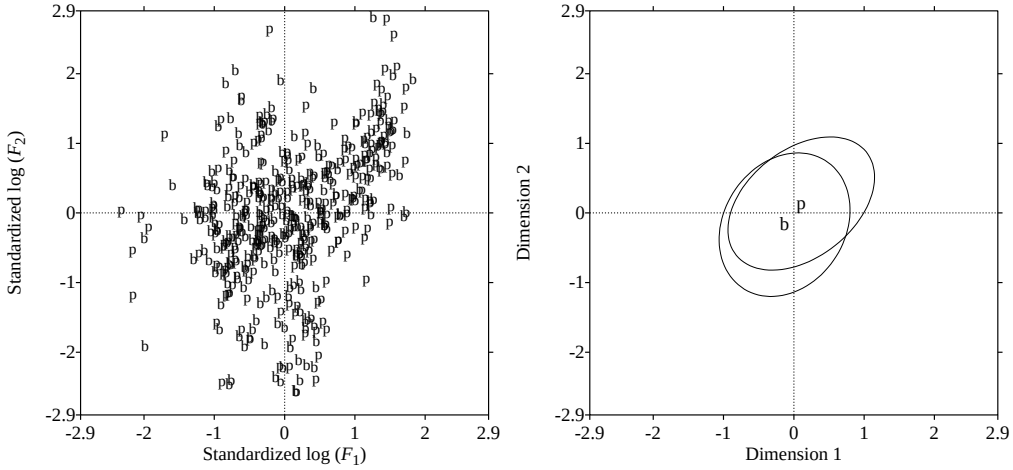

(a) Acoustic formant distribution

(b) Linear discriminant analysis

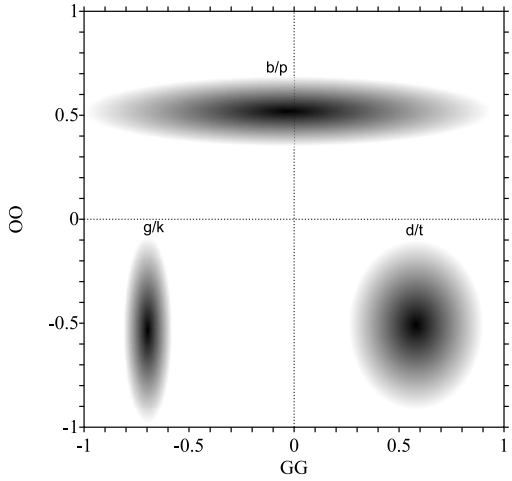
(c) Regional articulatory targets (RATs)

Figure 4.4: Acoustic and articulatory trajectories of the four primary cardinal vowels: front-low-rounded [ɑ], front-high-unrounded: [i], back-low-rounded: [ɒ], and back-high-rounded: [u].

to increase as the place of constriction moves backward from the lips to the soft palate, i.e., [b/p] at 19.1 ms, [d/t] at 33.0 ms, and [g/k] at 45.2 ms. In fact, it has been shown that human objects tend to perceive the speech sound in a decision-like *yes* or *no* manner, which is different from that of the other sensory organs such as the in-between decision on the continuously varying color spectrum




(a) Acoustic formant distribution of [b/p] (b) Linear discriminant analysis of [b/p]




(c) Regional articulatory targets (RATs)

Figure 4.5: Acoustic and articulatory trajectories of the three pairs of plosives, i.e., bilabial: [b/p], alveolar: [d/t], and velar: [g/k].

perceived by the visual system (Mottonen and Watkins, 2009).

The observation is also in line with the findings that the threshold of separation has the highest boundary value for velar, and the lowest for labial sounds (Niyogi and Ramesh, 2003). The cut-off VOT for syllable initial plosives also shows a 10.0 ms separation, e.g., the labial [b/p] at 30.0 ms, the alveolar [d/t] at 40.0 ms, and the velar [g/k] at 50.0 ms (Stouten and Hugo, 2009; Zue, 2004).

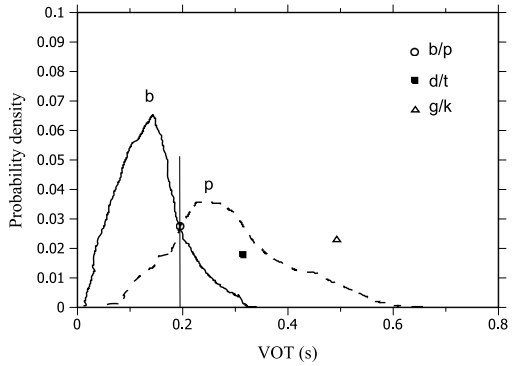



Figure 4.6: VOT distribution of [b/p] and the boundary points of the plosives in the recognition output.

These values are higher than the results reported here, but remain highly comparable, where the VOT boundary values and the separation rather indicate the difference in VTL of the speaker. This will be interesting for future studies.

## 4.5 Discussion

The initial hypothesis in this chapter is that the articulatory feature space presents a much smaller variance than the acoustic feature space. In the acoustic space, as shown in Fig. 4.4(a) and Fig. 4.5(a), two acoustic measures, formants: F1 and F2, are plotted for the consonants and vowels. In the articulatory space, as shown in Fig. 4.4(c) and Fig. 4.5(c), the articulatory trajectories, MVs: HG and GG, are plotted for them. To verify the hypothesis, two cardinal vowels, [ɑ] and [ɒ], each uttered 100 times by 5 speakers are extracted from the SCRIBE-TIMIT database. The entropy measure, $H_i(\mu, \sigma)$, for the two formant feature vectors are $[F1(5.93, 1.60), F2(10.14, 2.18)]$ for [ɑ], and $[F1(4.05, 0.16), F2(7.85, 3.02)]$ for [ɒ], all units in bark. In continuous speech, [ɑ] is often in a reduced form with a relatively large F2 variance that causes mis-classifications evidenced by the overlaps in Fig. 4.4(b). In the articulatory space, two MVs, GG and SG, are actually the principle controlling parameters for the tongue body forwarding and back-raising motions that distinguish [ɑ] from [ɒ]. Through iterative learning, the RAT is computed as a regional distribution, which reduces the acoustic variance within

each phone class (4.8). The center point of the RAT: $(GG_\mu, HG_\mu)$, is located at the central-front position for [ɑ]:(0.55, 0.30), and at a back position for [ɒ]: (-0.51, 0.52), as shown in Fig. 4.4(c). The resulting articulatory space is more compact than the acoustic space.

Furthermore, simulation results on the CV patterns suggest that though VOT is critical in the articulatory space, the acoustic realizations such as the energy burst in the consonantal onset and formant transitions in the vowel nucleus may or may not be fully audible in the acoustic output. Ragnier and Allen (2008) and Allen (2008) have previously discovered that the perception of [t] is entirely due to a single short ($\approx 20ms$) burst of energy, between 4 and 8 kHz. Furthermore, Li et al. (2010) have also found in their experiments that that consonants with similar events tend to form a confusion group. For example, [bɑ] and [vɑ] are highly confusable with each other because they share a common F2 transition, which is strong evidence that the auditory events, not articulatory features alone, also serve as the basic units for speech perception.

Close examination of the plosives-vowel patterns in the articulatory and the acoustic spaces also indicates that the identified bursts generally shift up in frequency for high vowels such as [i] but change little for low vowels such as [u], as suggested in (Li et al., 2010). The observation partly explains the fact that human speakers tend to maximize the categorical contrast between similar phonemes during the phonological encoding of CV syllabus, whereas the consonantal effects on the following vowels serve to ensure smoothness and continuity of natural speech (Martin and Jurasfsky, 2008). Furthermore, it shows that the multi-dimensional phonetic representations offers an alternative to model speech dynamics by utilizing the auditory and acoustic features for speech recognition and synthesis.

Though much remains unknown in the study of human sensory systems and mental states (Chomsky, 2006), the discovery of mirroring neurons has nonetheless suggested the bi-directionality of human speech production and perception that urges the collective investigation of both aspects to further advance the machine based speech recognition systems (Levelt, 1999). Previously Guenther et al. (2006) have constructed a neural model that attempted to organize the accumulated pool of articulatory and auditory data in a framework of humanoid sensory blocks which are analogous to the brain. Following their footsteps, Kröger et al.

(2009) have also built a neural model to enable parallel production and perception of simple syllables such as vowel clusters and CV patterns. However, up till now, the auditory and the articulatory features have mostly been used either as additional input streams or as internal representations in conventional ASR systems such as hidden Markov models (HMMs) (Siniscalchi and Lee, 2009), multi-layer perceptrons (MLPs) (Kirchhoff et al., 2002), time-delay neural networks (TDNNs) (Schuster and Paliwal, 1997), RBF-NNs (Yousefian et al., 2008), dynamic Bayesian networks (DBNs) (Frankel et al., 2007), and various hybrid paradigms (King et al., 2007; Trentin and Gori, 2001). The main difficulty in their ASR deployment is the non-linearities of the correlated articulatory and auditory features, which could result in undesirable feature redundancy as well as increased computational cost. For instance, many articulatory configurations could produce the same acoustic output, i.e., the *many-to-one* problem (Perrier and Ostry, 1996). On the other hand, human perception of speech is somewhat categorical, which projects the highly variant acoustic signals onto a limited set of meaningful phonemes (Chomsky, 2006). These issues will be addressed in the next chapter.

## 4.6 Summary

The articulatory trajectories and the acoustic qualities are two important aspects of the human speech. The proposed pronunciation modeling method roots in the concept that the highly variable and dynamic speech pattern can be decorrelated by the parallel speech production/perception mechanism, i.e., the motor representation in the articulatory channel and the perceptual cues in the auditory channel (Davis and Johnsrude, 2007; Scott and Johnsrude, 2003). This chapter clarifies the correlation of the parallel articulatory and auditory cues in natural English speech. It emphasize on the multi-attribute modeling of English pronunciations. The simulation study examines the absolute and the relative acoustic variance caused by different articulator gestures using the CV patterns. It verifies the hypothesis that the articulatory feature space is more compact than the acoustic feature space. The proposed multi-dimensional pronunciation modeling method

has several salient properties, such as the explicit control of the vocal apparatus movements. For instance, the bio-mechanical speech synthesizer mimics the human experience of speech acquisition. In addition, the articulatory-acoustic cues are mapped by two heuristic learning rules: the listener-oriented categorical speech perception and the speaker-oriented articulatory target approximation. Compared to conventional methods, the articulatory-acoustic pairings are more objectively distributed. This is beneficial for acoustic-articulatory inversion in the next chapter.

# Chapter 5

# Articulatory Phonetic Inversion for Improved Speech Recognition

## 5.1 Overview

This chapter investigates the use of production knowledge in automatic speech recognition[1]. It answers two key questions: how the articulatory data can be retrieved from the acoustic speech signal, and how the data can be used to improve the accuracy and the robustness of the recognition system. First, a synthetic speech dataset is constructed using the articulatory synthesizer. Next a clustering scheme is proposed too prepare the data for acoustic-to-articulatory mapping. Then a neural inversion module is implemented to retrieve the articulatory phonetic features from the acoustic features. Finally the inversion module is extended to a phoneme recognizer and test its performance in two diverse conditions: with different speakers and in noisy environments. Experiments show that the module preserves the articulatory phonetic details and obtains improved phoneme recognition performance than the acoustic hidden Markov baseline.

The chapter is arranged as follows. Section 5.2 gives the rationale on using production knowledge in automatic speech recognition (ASR). Section 5.3 explains the steps to obtain the articulatory-acoustic data through a clustering scheme. Section 5.4 implements the proposed neural model. Section 5.5 gives

[1]The manuscript was revised and re-submitted to Speech Communication in August 2012. Earlier results of the research were presented and published in (Huang and Er, 2011, 2012a).

the experiments and summarizes the results on speech inversion and on phoneme recognition. Section 5.6 discusses the experimental results. The chapter concludes in Section 5.7.

## 5.2 Speech Production Knowledge

The use of speech production knowledge can enhance the performance of ASR systems. For instance, the articulatory dynamics describe the smooth and continuous movements in the vocal tract, which induce the acoustic variabilities of human speech. Compared to the acoustic cues, the articulatory features/cues/dynamics are slow varying and are constrained by the physiological property of the speech apparatus. However, the articulatory dynamics are usually unknown or not readily available in ASR applications. Many articulatory based processing methods have been proposed to model the acoustic-phonetic variations in a number of frame-based, segment-based, and acoustic landmark systems (King et al., 2007; Stevens, 2002). One such method derives the phonological articulatory features(PAFs) from the phonological rules of speech (e.g., manner, place of articulation, and voicing). The PAFs have improved ASR performance over the acoustic hidden Markov model (HMM) baseline (Frankel et al., 2007; Kirchhoff et al., 2002; Saenko et al., 2005; Scharenborg et al., 2007). Moreover, direct articulatory data have been collected using electromagnetic articulograph (EMA), X-ray analysis, and laryngograph of the human speakers (Richmond, 2009). In theory, these data are more accurate than the PAFs in describing the dynamics of the vocal apparatus. Therefore, many inversion techniques have been proposed to estimate the articulatory data from the readily available acoustic recordings (e.g., using microphone).

Mapping the articulatory features with the acoustic features requires a speech corpus with parallel acoustic and articulatory data. One option is to use the direct articulatory data, such as the EMA recordings in the multichannel articulatory (MOCHA) corpus (Wrench, 1999). For example, Richmond (2009) trained a mixture density neural network to obtain accurate flesh-point articulatory trajectories from the acoustic spectral data. However, the author pointed out that the network performance degraded due to the inconsistencies in the recordings, which

were caused by the physiological differences of the speakers (Richmond, 2009). In addition, the EMA corpus usually has a limited number of uniformly distributed phonetic events such as diphthongs and dialectics. This makes training difficult in the acoustic based HMM recognizers especially for continuous speech. Another option is to construct an articulatory codebook. For example, Mitra et al. (2011) generated a synthetic database containing the articulatory-acoustic data pairs using an articulatory speech synthesizer. The method has also obtained good performance for a series of acoustic-to-articulatory inversion experiments (King et al., 2007; Schroeder, 2004; Schroeter and Sondhi, 1994). One advantage of the codebook method is that the new data can be generated easily by the synthesizer at any time. This is especially useful when dealing with unfamiliar speech events such as out-of-vocabulary words and phonetic variations. Furthermore, the physiological parameters such as the lung pressure and the vocal tract shape can be explicitly defined in the synthesizer, which renders the generated codebook more consistent than the EMA recordings (Boersma, 1998; Merhav and Lee, 1993). This study uses an articulatory synthesizer developed by Mermelstein (1973) to prepare a synthetic dataset. Moreover, the neural based articulatory phonetic inversion (API) model addresses the non-linearity issue in inversion and recognition experiments. The API model includes two concatenated neural modules as shown in Fig. 5.1. The inversion module uses Elman's recurrent networks (RNN) to estimate the articulatory parameters. It is initially trained on a synthetic dataset to estimate the articulatory phonetic features (APFs) from the acoustic speech signal. To deal with the non-uniqueness in the articulatory-acoustic data pairs, the clustering scheme is designed to find the minimum set of acoustic and articulatory classes. The estimated APFs are smoothed by a low pass filter to eliminate unrealistic articulatory movements. The second module uses the feed-forward neural network to classify the phonemes from the estimated articulatory parameters. It is trained on a natural speech corpus to map the smoothed APFs to phones, and to phonemes, which are the abstract linguistic units of the English speech.

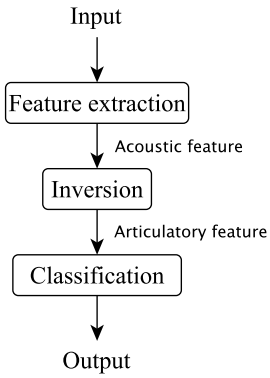



Figure 5.1: Block diagram of the articulatory phonetic inversion model.

## 5.3 Data Acquisition

### 5.3.1 Parametrization

There are two methods of obtaining the articulatory-acoustic data pairs from the articulatory synthesizer (King et al., 2007; Schroeter and Sondhi, 1994). The first is to define the maximum and the minimum values of the APF parameters, and fill the bounded region between the extreme positions (Guenther et al., 2006). The other is to randomly sample the whole articulatory space with all the possible configurations of the APF parameters (Perrier et al., 2005). The first method only generates the realistic tract shapes. The second method gives a fair coverage of all the possible tract shapes in the articulatory space, but the processing time is often high, where Mitra et al. (2011) reported that it would take 85 seconds to generate a mono-syllabic word, e.g., the word "one".

Since the phonetic event occupies a spread region, rather than isolated points, in the articulatory-acoustic space (Damper and Harnad, 2000; Kielar et al., 2011; Mottonen and Watkins, 2009), the first bounded region method is used to generate the required data. The midsagittal view of the vocal tract in Mermelstein's synthesizer is drawn again in Fig. 5.2, produced using the PRAAT software (Boersma, 1998). This study focuses on measuring the articulators which are critical for human speech production. The set of 13 APFs include the relative position of the tongue root $(TR_x, TR_y)$, the tongue body $(TB_x, TB_y)$, the tongue

tip $(TT_x, TT_y)$, the lower lip $(LL_x, LL_y)$, the upper lip $(UL_x, UL_y)$, the glottis width (GW), the jaw angle (JA), and the velum opening (VO). The tract variables (TVs) are parameters used in the articulatory synthesis experiments, and APFs are parameters used in the speech recognition experiments. Different from the TVs, APFs include the glottis and the velum settings (cf. Section 3.3 in Chapter 3 and Section 4.2.1 in Chapter 4). Each APF represents one articulatory dimension. The synthetic data is generated in two steps. First, the extreme geometric shapes of the vocal tract are defined. The boundary or maximum APF values are based on the observations of speech synthesis experiments and EMA measures. Except for JA which is in radian (*rad*), the others are in *mm*. Second, the target regions are randomly sampled into uniformly distributed pallet points, the regional articulatory targets (RATs). The method prevents the articulatory synthesizer from generating unrealistic vocal tract shapes, and it gives a fair coverage of all the possible tract shapes without processing the whole articulatory space. There are totally seven articulatory target regions for the English phonemes, as shown in Table 5.1.

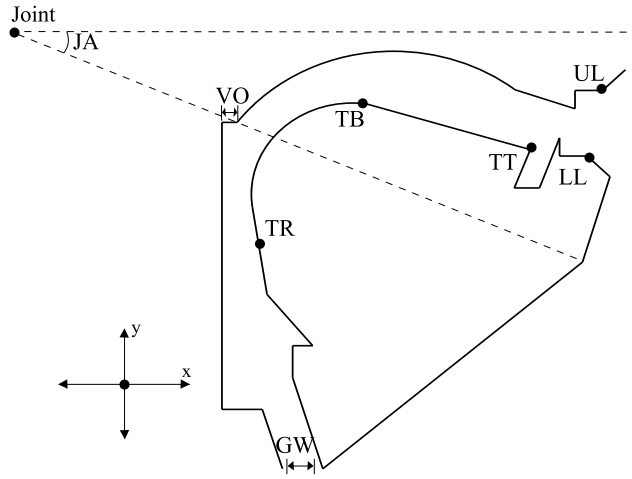



Figure 5.2: Vocal tract geometry and the pallet positions of the APFs.

During simulation the articulator moves from a neutral/resting position or one RAT point to another RAT point. The articulatory synthesizer uses a normal speaking rate with 50 ms silence at the beginning and the end of each utterance (Mermelstein, 1973). The pitch range is between 60 Hz (male) to 300 Hz (female),

Table 5.1: Articulatory target regions and the 45 English phonemes.

| Region | APF | Boundary (mm) | Phonemes |
|---|---|---|---|
| Lips | $UL_x$ | 20 | p, b, v, f, m, w, ow, oy, uh, uw; |
| | $UL_y$ | 15 | |
| | $LL_x$ | 20 | |
| | $LL_y$ | 15 | |
| Tongue tip | $TT_x$ | 20 | ch, sh, th, dh, s, z, t, d, dx, l, r, y; |
| | $TT_y$ | 20 | |
| Tongue body | $TB_x$ | 12 | sh, r, aa, ae, ah, aw, ay, eh, er, ey, ih, iy, ow, oy, uh, uw; |
| | $TB_y$ | 18 | |
| Tongue root | $TR_x$ | 10 | k, g, ng, er; |
| | $TR_y$ | 10 | |
| Glottis | GW | 2 | b, d, g, z, v, dh, aa, ae, ah, aw, ay, eh, er, ey, ih, iy, ow, oy, uh, uw; |
| Velum | VO | 5 | m, n, ng |
| Jaw | JA | 1.2 (rad) | aa, ae, ah, aw, ay, eh, er, ey, ih, iy, ow, oy, uh, uw; |
| Silence | | | bcl, pcl, dcl, tcl, gcl, kcl, sil (pau). |

which is typical for human natural speech. For monophthongs, the transition time from the rest position to the selected RAT is 30 ms. For diphthongs and longer syllables, the transition from one RAT to another is 20 ms. For example, in consonant-vowel production the transition from constriction to total release is 30 ms followed by the static vowel articulation. The position of the articulators in the transition region changes stepwise from one time slice (5 ms) to the next. Similar settings were also used by Perrier et al. (2003) in their bio-mechanical tongue model for simulation of consonant-vowel sequences. The acoustic signal is processed by Bark-scale filter bank to produce the auditory spectra (Brown and Cooke, 1994; Gajic and Paliwal, 2006). The filter bank consists of 23 overlapping bandpass Gamma-tone filters each with a bandwidth of 3 Bark. The center frequencies of the filters are uniformly distributed on the Bark scale between 100

and 4000 Hz. For comparison with the standard mel-frequency cepstral coefficients (MFCCs), an analysis window (Hamming) of 15 ms with a frame shift of 10 ms is used, which matches the sampling rate of the articulatory parameters. Principle component analysis (PCA) is used to de-correlate the sub-band spectral energy vectors and reduce the dimensionality, which produces the static cepstral coefficients, bark-frequency cepstral coefficients (BFCCs). The first 12 discrete cosine transform coefficients are augmented by their first and the second order derivatives. The derivatives (or delta) coefficients are calculated using two frames of past context and two frames of future context. Log energy is attached resulting in 39 dimensional BFCCs. The articulator remains in the final RAT for 100 ms. The sampling rate for the articulatory parameters is 100 Hz. The sampling rate for the acoustic waveform is 8 kHz.

### 5.3.2 Data Clustering

The data acquired is the synthetic speech. In the synthetic dataset, there are initially seven articulatory regions of APF vectors, excluding the silence region. The main difficulty is that there may be more than one vocal tract shapes in the articulatory space that generate similar acoustic signals, known as the the non-uniqueness of speech inversion (Bickford, 2006; Toda et al., 2008). However, Sondhi and Schroeter (1987) observed that though non-unique in a single global map, the acoustic-to-articulatory mapping is locally unique after partitioning based on their acoustic and articulatory similarities. In this study, the RATs of the articulatory feature vectors are uniformly distributed with equal distance in the articulatory space, so the clustering scheme aims to further split the synthetic data-pairs into subclusters based on their acoustic similarities. The clustering scheme was previously used to reduce the sparsity of features in face recognition experiments (Er et al., 2005).

Fig. 5.3 illustrates the proposed clustering scheme, where the uniformly distributed RATs along two articulatory dimensions and the corresponding acoustic classes are shown in (a) and (b) respectively. The scheme aims to find an optimal $N_B$ acoustic subclusters based on the Euclidean distance of the feature vectors. Each of the subclusters is located in a scope with a controllable mean radius

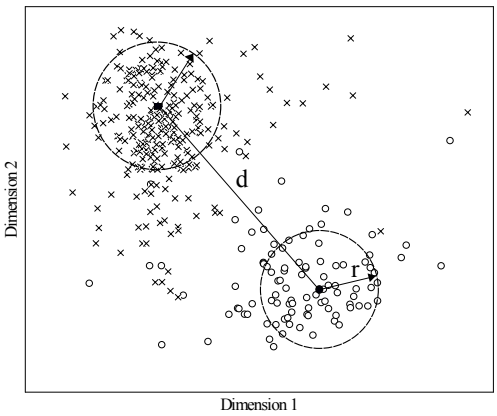



Figure 5.3: Illustration of the proposed clustering scheme.

of $\gamma$. The scheme prevents the feature vectors with large variations from being grouped in the same subcluster. Within each subcluster the acoustic vectors are close to each other, and the articulatory vectors are also close to each other. If one acoustic vector (BFCCs), denoted as $B$, can map to multiple articulatory vectors (APFs), denoted as $A$, in at most $N_A$ articulatory clusters, the initial synthetic dataset can be split into a total number of $N_A \times N_B$ subclusters of data pairs $(A, B)$. The clustering algorithm is as follows:

1. For the $a^{th}$ articulatory region, $a = 1, 2, \cdots, n_a$, where $n_a = 7$ is the number of articulatory regions. Let $n_b$ be number of clusters in the acoustic space, initially $n_b = n_a$

2. Find two reference RAT points $A_i^a, A_j^a$ from the $a^{th}$ articulatory region, which have the largest Euclidean distance $d_{ij}^a$ in the region.

3. Find the two reference acoustic vectors $B_i^a, B_j^a$ from the corresponding acoustic cluster.

4. Compute the acoustic distance from the two reference vectors to the other vectors in the acoustic cluster, $B_k^a$, and $k = 1, 2, \cdots, n_k$, where $n_k$ is the number of samples in the $a^{th}$ cluster, using the Euclidean distance in decibel

(dB):

$$d_{ik} = \frac{10}{ln10}\sqrt{\sum_{x=1}^{39}(B_{xi}^{a} - B_{xk}^{a})^2}, \tag{5.1}$$

$$d_{jk} = \frac{10}{ln10}\sqrt{\sum_{x=1}^{39}(B_{xj}^{a} - B_{xk}^{a})^2}. \tag{5.2}$$

5. Compute the mean and the standard deviation of the acoustic distance:

$$\mu_{d_{ik}} = \frac{1}{n_k}\sum_{k=1}^{n_k} d_{ik}, \tag{5.3}$$

$$\sigma_{d_{ik}} = \sqrt{\frac{1}{n_k}\sum_{k=1}^{n_k}(d_{ik} - \mu_{d_{ik}})^2} \tag{5.4}$$

$$\mu_{d_{jk}} = \frac{1}{n_k}\sum_{k=1}^{n_k} d_{jk}, \tag{5.5}$$

$$\sigma_{d_{jk}} = \sqrt{\frac{1}{n_k}\sum_{k=1}^{n_k}(d_{jk} - \mu_{d_{jk}})^2} \tag{5.6}$$

6. Define two circular scopes with centroids at $B_i^a$ and $B_j^a$ in the acoustic cluster, each with a radius of $\gamma_i^a$ and $\gamma_j^a$ respectively:

$$\gamma_i = \mu_{d_{ik}} + \alpha \cdot \sigma_{d_{ik}}, \tag{5.7}$$

$$\gamma_j = \mu_{d_{jk}} + \alpha \cdot \sigma_{d_{jk}}, \tag{5.8}$$

where $\alpha$ is a positive clustering factor initially set as 1, and its effect on the inversion process will be evaluated in the experiments.

7. The degree of overlapping between the two scopes depends on the data sparsity, and it has three outcomes:

   (a) If $d_{ij} \geq \gamma_i + \gamma_j$, the two scopes do not overlap, split the acoustic cluster into two subclusters, and update $n_b = n_b + 1$;

(b) ElseIf $d_{ij} \leq |\gamma_i - \gamma_j|$, the two scopes overlap greatly, and the cluster is tight;

(c) Else $|\gamma_i - \gamma_j| \leq d_{ij} \leq \gamma_i + \gamma_j$, the two scopes overlap, and they form an eclipse like data distribution; Define a new subcluster with a centroid at the mid-point between the two reference centroids, and with a radius of $0.5 d_{ij}$;

(d) EndIf.

8. Go to the next articulatory region; $a = a + 1$, and repeat step 1 - 7 for all regions.

9. In the resulting acoustic space, the random samples which do not fall into any of the subclusters are discarded from the synthetic dataset.

In the extreme cases, if the scope is too large, the training samples may be overly generalized, which results in too few distinctive clusters. If the scope is too small (e.g., $\alpha = 0$), the training samples may form too many clusters, which increases the computation cost of the mapping.

## 5.4 Neuron Model

The limitation of many current text-to-speech (TTS) and ASR systems are due to the fact that they are not faithfully designed with respect to the human neural processes of speech production and perception. Current ASR systems are subject to many constrictions such as the vocabulary size, the speaker, and the noise contamination (Scharenborg et al., 2007). Current TTS systems are also limited when dealing with the speaker characteristics an the prosodic naturalness (Martin and Jurasfsky, 2008). These limitations are the main areas that human easily outperforms the machine based systems. In the literature, a variety of brain imaging studies have clarified the role of different subcortical and cortical brain regions for speech production as well as for speech perception. Other studies have also shown the inter-correlation of the production and the perception pathway (Kröger et al., 2009). There are few functional neural models which explain the complex neural sensory-motor and the auditory processes of human speech

processing. So the proposed model aims to implement the two aspects together, which is capable of imitating human speech production and perception based on neuro-physiological and neuro-psychological knowledge of speech processing.

The proposed API model concatenates two modules: inversion and classification. The neural topology is shown in Fig. 5.4. Neural networks have shown good performance in previous speech inversion experiments, where the feed-forward connections are used to perform a non-linear mapping between the acoustic parameters and the articulatory parameters (Mitra et al., 2011; Toda et al., 2008). Here the inversion module implements the Elman RNN in which the hidden layer units are fed into the input layer (Elman, 1990). The feedback mechanism enables the RNN to maintain a short-term memory, or the prediction knowledge, from the input vectors. This is useful in capturing the correlations in the time-sequential feature vectors such as the dynamics of the articulatory parameters across the time frames (Schroeter and Sondhi, 1994).

In the inversion module, the RNN infers the articulatory features vectors, APFs, from the acoustic feature vectors, BFCCs. A low-pass filter with a cut-off frequency at 15 Hz is used to smooth the APFs by eliminating unrealistic or abrupt articulatory movements. The smoothed APFs are input to the second classification module, where the articulatory parameters are mapped to the phonemes. The classification module includes a phonetic layer and a phoneme layer, as shown in Fig. 5.4. The phoneme is a linguistic unit, which may have more than one acoustic realizations, or phones. One neuron in the phoneme layer links to several neurons in the phonetic layer. The connections are determined by the phonetic information in the Spoken Corpus Recordings in British English (SCRIBE)-Texas Instrument and Massachusetts Institute of Technology (TIMIT) corpus, where the broad *phoneme* annotations and the narrow *phonetic* annotations are given in parallel on the acoustic waveforms.

The API model has five layers. Layer 1 is the acoustic layer with 39 nodes, which correspond to the input BFCC with the first and the second derivatives. Layer 2 is the recurrent layer with 100 nodes. Layer 3 is the APF layer with 13 nodes, which correspond to the number of articulatory channels. Layer 4 is the phonetic layer with 255 nodes, which is the number of narrow phonetic annotations in the SCRIBE-TIMIT dataset. Layer 5 is the output phoneme layer

with 45 nodes, which represent the broad phoneme annotations. The activation function of the output layer is the softmax function,

$$f(x_i) = \frac{exp(x_i)}{\sum_{k=1}^{K} exp(x_k)}, \tag{5.9}$$

where K is the number of units in the output layer. The outputs are interpreted as phoneme probabilities, which sum to 1. The other layers use the tanh activation function.

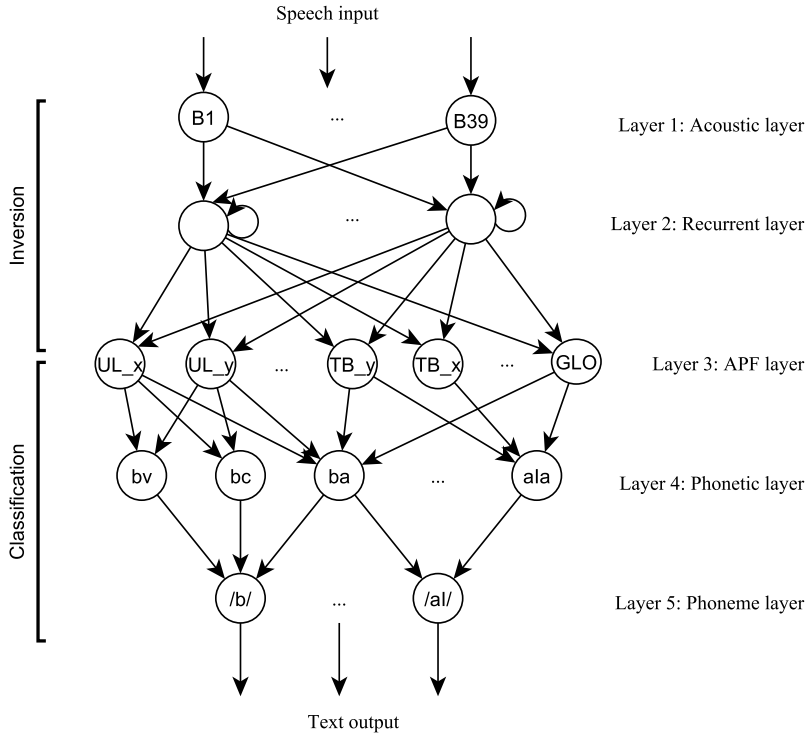



Figure 5.4: Neural topology of the API model.

## 5.5 Simulation

The API model is tested in the speech inversion and recognition experiments. Initially the synthetic speech dataset optimizes the parameters of the RNN based

inversion module. Then a natural speech dataset is used for the phoneme recognition tasks on the the API model. The natural dataset is extracted from the SCRIBE-TIMIT corpus (Huckvale, 2004), which consists of 100 phonetically rich TIMIT SX sentences repeated by 5 native English speakers, 4 males (MAC, MAE, MAF, MAM) and 1 female (FAA), all with the received pronunciation.

### 5.5.1 Acoustic HMM Baseline

The recognition performance of the API model is compared with that of an acoustic HMM baseline, implemented using Cambridge's HTK tools (Young et al., 2006). The HMM recognizer has 45 context independent monophone models with 5 states per phone and 15 diagonal covariance Gaussians per state. It also includes a 3-state silence model, which shares the middle state with a 1-state short pause model. The acoustic features in the HMM baseline are the standard 39-dimensional MFCCs including the log energy, the 12 static cepstral coefficients, and the appended first and second derivatives. Different from the Bark scale filters, the acoustic baseline has 23 triangular filters which are uniformly distributed on the mel scale between 100 and 3000 Hz with 50% overlap. A pre-emphasis filter with coefficient 0.97, and a 256 point fast Fourier transform (FFT) are used. Other settings for the MFCCs are the same as the BFCCs.

### 5.5.2 APF Inversion on Synthetic Speech

This study report the root mean square error (RMSE) of the estimated articulatory parameters, APFs, for the RNN-based inversion module on the synthetic dataset. It divides the synthetic dataset of the articulatory-acoustic data pairs into five partitions. Each of the partitions is used in turn for testing, while the other four are used for training. During the five training sessions, the inversion module minimizes the RMSE. The clustering scheme defines the scope of each subcluster using $\gamma = \mu + \alpha \times \sigma$, which can be interpreted as the acoustic distance among the training samples. Fig. 5.5 plots the averaged RMSE of the APFs with different $\alpha$ values. The clustering factor $\alpha$ shows low RMSE between 1.0 and 1.5. For the following experiments $\alpha = 1.2$ is used, which creates about 170 subclusters in the synthetic data.

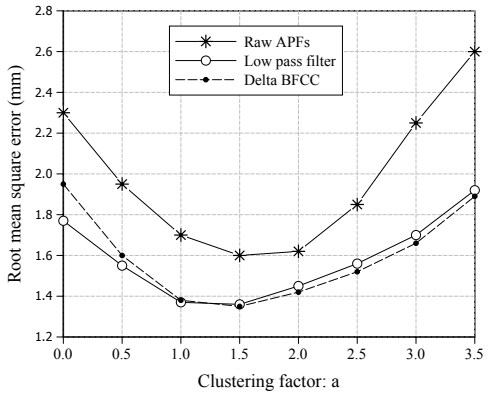



Figure 5.5: Averaged RMSE of the APFs with different clustering factor: $\alpha$ on the synthetic dataset.

Table 5.2 shows the RMSE results. Here jaw height (JH) is calculated as the distance from the joint to the upper incisor times the tangent of the jaw angle. The low-pass filter can effectively reduce the RMSE of the APFs (5% significance level using t-testing). Toda et al. (2008) have also obtained reduced error rate using the low-pass filtering in their speech inversion experiments. The delta BFCCs do not obtain higher accuracies than the low-pass filter. It may be due to the fact that the smoothed APFs already account for some amount of the contextual variations in the acoustic input, and the delta coefficients do not introduce new knowledge sources. In the articulatory space, the RMSE also shows different characteristics in each APF dimension, as shown in Table 5.2. For example, the low-pass filter obtains higher RMSE reduction for the $TT_y$ and the JH than the others. The relatively high error reduction rates of the $TT_y$ and the JH values are mainly due to their large dynamic movements in their articulatory regions for many consonant-vowel syllables. For instance, the tongue tip is involved for the production of dental plosives (e.g., /t, d/) and the jaw movement is involved during the articulatory synthesis of the low vowels (e.g., /aa, ow/).

Table 5.2: RMSE of the APFs in the inversion experiment.

| | RMSE (mm) | | |
|---|---|---|---|
| | Un-smoothed | Smoothed | Reduction (%) |
| $UL_x$ | 0.91 | 0.77 | 15.4 |
| $UL_y$ | 1.13 | 0.99 | 12.4 |
| $LL_x$ | 0.81 | 0.72 | 11.1 |
| $LL_y$ | 1.15 | 1.10 | 4.3 |
| $TT_x$ | 2.09 | 1.98 | 5.3 |
| $TT_y$ | 2.69 | 2.15 | **20.1** |
| $TB_x$ | 2.15 | 1.93 | 10.2 |
| $TB_y$ | 2.12 | 1.88 | 11.3 |
| $TR_x$ | 2.13 | 1.85 | 13.1 |
| $TR_y$ | 2.22 | 1.92 | 13.5 |
| GW | 0.25 | 0.21 | 16.0 |
| VO | 0.54 | 0.51 | 5.6 |
| JH | 2.55 | 1.96 | **23.1** |
| Average | 1.60 | 1.38 | 13.4 |

### 5.5.3 Phoneme Recognition on SCRIBE-TIMIT

When using the API model for the phoneme recognition task, it requires two stages of training. First, the inversion network is updated on the synthetic speech dataset, called the initialization stage. Second, the classification module is updated on the natural speech dataset, called the bootstrapping stage. Unlike the inversion task, the API model optimizes the cross-entropy instead of RMSE, where the phoneme recognition rate is reported. For the natural speech, the SCRIBE-TIMIT corpus provides detailed acoustic-phonetic information, which balances the trade-off between the perceptual invariance of the human listeners and the articulatory effort of the native speakers on the set of phonetically rich SX sentences. It also denotes the minimum audible changes across the phonetic annotations as perceived by the linguists in the acoustic speech signal. The articulatory-acoustic data in the synthetic dataset tunes the inversion module, where the acoustic-phonetic data in the natural dataset tunes the classification

module. The proposed API system is trained in clean conditions and tested for the speaker and the noisy testing.

#### 5.5.3.1 Speaker Testing

The recognition accuracy is the percentage that the frames are correctly classified, i.e., the frame error rate (FER), which are summarized in Table 5.3. The inside testing uses the same speakers as from the training set. 80% of the sentences for each speaker are used for training, and the rest 20% for testing. The outside testing uses different speakers for the training and the testing. Each of the 5 speakers is used for testing, where the other 4 are used for training. The reported FER is an average of the 5 test runs.

Table 5.3: Frame level accuracy (%) of the speaker independence testing.

| | Inside testing | | Outside testing | |
|---|---|---|---|---|
| | Best-1 | Best-3 | Best-1 | Best-3 |
| HMM | 82.6 | 85.4 | 69.2 | 80.1 |
| API | 88.5 | **92.3** | 74.5 | **83.0** |

This study compares the phoneme recognition accuracy of the API model on the SCRIBE-TIMIT data with the acoustic HMM baseline. For both inside and outside testing, the API model outperforms the acoustic-HMM baseline (5% significance level). It is possible that the contextual variations in the acoustic input are represented as the articulatory dynamics in the RNN. In addition, the phonetic layer in the neural structure explicitly maps the dialectal variations with the phonemes. The observation agrees with previous claim that the use of AFs can improve the ASR performance. Different from the phonologically derived PAFs, the English phonemes are represented as distributed RATs in the articulatory space (Frankel et al., 2007; Scharenborg et al., 2007). In the literature, phoneme recognition accuracy as high as 75.6% was reported on the TIMIT SX sentences using the tandem multi-layer perceptrons (MLPs) with an adjustable neural structure, where the number of neurons in the hidden layer of neural networks was increased until the error rate saturated (Schwarz et al., 2006).

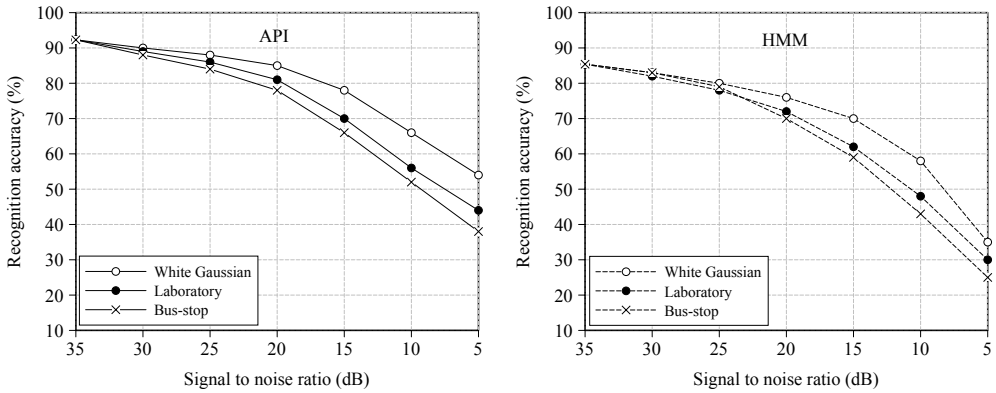
#### 5.5.3.2 Noisy Testing

This section reports the phoneme recognition performance of the API model in three noisy settings: the white Gaussian noise, the laboratory noise, and the bus-stop noise. The laboratory and the bus-stop noise are from the corpora in (Nasibov and Kinnunen, 2012). They are selected to evaluate the adaptability of the API model in practical applications. The laboratory has several low frequency noise sources including the computer working stations, the air-conditioners, and the background noises inside the lab. The bus-stop has noise sources from the the passengers, the passing cars, and the engine noise of the buses. In the natural dataset, 80% of the sentences for each speaker were used for training, and the rest 20% for testing, which contains 186,342 frames in total.

This study compares the robustness of the proposed method with the cepstral mean subtraction technique in the MFCC-HMM baseline, where the mean value of the MFCCs across the input time frames is calculated and then subtracted from each frame. Fig. 5.6 shows the recognition accuracy of the API and the HMM recognizers in noisy testing conditions. The phoneme accuracy decreases with increasing noise levels. The bus-stop noise introduces the highest deduction. The average accuracy in noisy testing is 74.8 % for the API recognizer, and 66.2 % for the HMM recognizer. The performance improvement confirms that the articulatory features are more robust against noise contaminations compared to the acoustic features (Frankel et al., 2007; King et al., 2007; Kirchhoff et al., 2002). In fact, the APFs represent the underlying articulatory movements which are less error prone in the presence of noise for ASR. Yet the APFs are inverted from the clean synthetic corpus, which introduces mis-match between the training and testing conditions.

### 5.5.4 Phoneme Recognition on TIMIT

To test the hypothesis that the API module can deal with the pronunciation variations of natural speech data, the original TIMIT dataset is used in the second recognition experiment. The available 450 SX sentences produced by 630 speakers (192 female/438 male), 5 sentences per speaker, from all dialect regions (DR1 to DR8) were used to test the accuracy of the API model on unfamiliar speech




(a) API (b) HMM

Figure 5.6: Recognition accuracy of the API and the HMM recognizers in increasing noisy levels.

events, and no sentence text appeared in both the training and the testing sets. The testing set contains a total of 168 speakers and 840 utterances, accounting for about 26.7% of the total speech material. The training set contains 462 speakers and 2310 utterances, accounting for about 73.3% of the total speech material. During training, the API classifier is monitored on the 100 validation sentences, and terminated the process when the global error of back-propagation on the validation dataset approached static to avoid over-fitting.

Besides the acoustic HMM baseline, this study also implements two types of artificial neural networks (ANNs) based recognizers, MLP and RNN, with comparative configurations as in the API model. The MLP classifier consists of three layers: one input layer with 390 nodes for 10 frames of 39-dimensional MFCCs, one hidden layer of 300 nodes each with a tanh activation functions, and one output layer with softmax activation function and 61 nodes. The RNN classifier has the same settings as the MLP at the hidden and the output layer, except that it has a time delay of one time frame at the input layer, which allows the network to infer knowledge about temporal dynamics from the input feature streams (Schuster and Paliwal, 1997; Strom, 1997). Similar to the API model, the ANN based classifiers used the cross-entropy objective function as the optimization criterion. All networks run 55 cycles of the back-propagation algorithm with a momentum of 0.7 and a gain of $1 \times 10^{-7}$.

Table 5.4 summarizes the classification results of the seven phoneme recognizers on the TIMIT testing set. For the best-1 FER, the correct label is ranked the highest in the decoded frame. For the best-3 FER, the correct label is among the top three labels in the results.

Table 5.4: Frame level recognition results of the different stream lines on the TIMIT testing set.

| | FER: Best-1 (%) | FER: Best-3 (%) |
|---|---|---|
| MFCC + HMM (CI monophone) | 48.0 | 60.4 |
| MFCC + HMM (CD tri-phone) | 51.9 | 62.3 |
| MFCC + MLP (10 frames) | 62.0 | 77.1 |
| MFCC + RNN (1 delay) | 68.9 | 83.1 |
| BFCC + RNN (1 delay) | 71.6 | 85.5 |
| BFCC + API (PAF) | 69.0 | 85.1 |
| BFCC + API (APF) | **75.0** | **85.6** |

Using the same front end, MFCCs, the four back ends, CI-HMM, CD-HMM, MLP, and RNN obtained increasing level of accuracy, where RNN outperformed the rest for both best-1 and best-3 ranking. Similarly, using the same back end, RNN with 1 time frame delay, the two front ends, MFCC and BFCC, achieved comparable accuracy, where BFCC gained slightly higher accuracy than MFCC with 2.7% improvement for best-1 ranking and 2.4% for best-3 ranking. The average improvement obtained by BFCCs over MFCCs is less than 3%, which is not as high as that obtained by the RNN over other back ends, where it outperformed the CI-HMM by 20.9%, the CD-HMM by 17.0%, and the MLP by 6.9% in the best-1 ranking. This observation shows that the MFCC features are still in many ways a near optimal representations in current ASR systems, and that there are more to be gained at the back end using more sophisticated classification methods.

At the frame level, the use of AFs provides another interesting observation. The best-1 measures of the two articulatory feature (AF) streams in the API model show higher FER than the others, and this advantage persists in the best-3 measures compared with the purely acoustic-phonetic cues. However, the im-

provement obtained by API model over the RNN model is not absolute, where PAFs show slightly lower accuracy than the RNNs. The strength of RNN extends to its utmost in the API module using APFs, demonstrating higher accuracy in the best-1 and best-3 results than the PAFs, which suggests that that the continuously valued APFs are more adequate for pronunciation modeling than the quantized PAFs, especially when the phonetic segments contain a significant amount of variations as in the SX sentences.

Table 5.5 summarizes the performance of the recognition modules as a function of signal-to-noise ratio (SNR) on the TIMIT database. Phone recognition accuracy are measured against SNR from 0 to 30 dB (clean speech) using artificially added white Gaussian noise. It appears that not only does the API module lower the PER, it also shows robustness against noise contamination. The API model with the APFs sustained its advantage of higher accuracy in noisy settings, and the performance degradation rate is 0.95% per dB, which is slightly lower than the others. It is highly probable that the individual AFs, PAFs or APFs, deteriorates less strongly than the other acoustic based classifiers, resulting in the robust performance in adverse conditions (Kirchhoff et al., 2002). Moreover, the use of the auditory based BFCCs in the API model also suppresses noise contamination by discarding irrelevant acoustic cues in the signal. And the multi-dimensional phonetic cues in the API model obtained the best performance in noisy testings. In Table 5.5, the auditory based BFCCs only shows slightly higher performance than the MFCCs in the same RNN classifier. Yet the major contribution of the API superiority comes from the proposed pronunciation models, which are more accurate and more reliable for articulatory inversion.

## 5.6 Discussion

High performance phoneme recognition is a challenging task for machine based speech recognizers. This chapter examines the feasibility of using multiple knowledge sources, i.e., the articulatory and the auditory features, to improve speech recognition performances. It implements the neural based API model with the concatenated RNN and MLP structure to realize simple yet elegant speech recognition. This structure elevates the problem of temporal representations in speech

Table 5.5: The effect of noise contamination on phone recognition accuracy (%) as a function of SNR (dB).

| | 30 | 20 | 10 | 0 | degradation (%)/dB |
|---|---|---|---|---|---|
| MFCC + HMM (CD tri-phone) | 51.9 | 47.8 | 41.5 | 20.7 | 1.04 |
| MFCC + MLP (10 frames) | 62.0 | 48.5 | 45.3 | 29.4 | 1.09 |
| BFCC + RNN (1 delay) | 71.6 | 67.2 | 61.9 | 39.8 | 1.06 |
| BFCC + API (PAF) | 69.0 | 67.8 | 60.6 | 38.0 | 1.03 |
| BFCC + API (APF) | 75.0 | 71.5 | 63.3 | 46.4 | **0.95** |

patterns, and it took advantage of the posterior probabilistic estimation by MLPs at the frame level (Martin and Jurasfsky, 2008).

However, there remain a few issues in the current system. One issue is the physiological specifications in the synthesizer, where the intrinsic properties of the speaker are possibly over-generalized. For example, the physiological matrices $M$, $K$, & $B$ in the tongue muscles require sophisticated modeling methods for realistic approximation (Birkholz et al., 2011; Perrier and Ostry, 1996). Thus it will be interesting to use the available articulatory recordings, e.g., the MOCHA-TIMIT corpus, to monitor the articulatory-acoustic mapping in the computational model (Wrench, 1999). Another issue is the use of higher level knowledge such as semantics and syntactics in the neural model for recognition of complete word or phoneme strings. Instead of $Pr(O_i|G_k)$, the network output will need to estimate the continuous probability, $Pr(O_1, O_2, O_2, \cdots, O_I|G_k)$, with consideration of the contextual constraints.

### 5.6.1 Phoneme Recognition Accuracy

The results in this study agree with previous studies on the use of speech production knowledge in ASR systems. Table 5.6 compares the phoneme recognition accuracy of the API model with the existing recognizers on the TIMIT sentences. Previously Jeon and Juang (2007) have used a set of 12 dimensional cortical response patterns derived from a central auditory model in a HMM based recognizer, which obtained lower phoneme accuracy than the MFCC baseline in clean

testing. The system acquired certain degree of robustness in noisy conditions with slower performance degradation over the MFCC baseline. Other auditory inspired methods have also been successfully employed to exploit the use of auditory features such as the adaption model proposed by Holmberg et al. (2006), which obtained 46% relative word error reduction for clean speech training on the AURORA 2 recognition task. Siniscalchi and Lee (2009) used a set of MLP classifiers to extract the phonological AFs in a hybrid HMM/ANN recognition system, where phoneme recognition accuracy on the TIMIT SX and SI sentences was improved by 5.3% for clean testing. King and Taylor (2000) obtained 63.3% phoneme accuracy using RNNs and AFs on the TIMIT sentences. Schuster and Paliwal (1997) proposed a bidirectional recurrent neural network, which integrated the forward RNN and the backward RNN to take into account all the features available from the input frames in both directions. They obtained similar phoneme recognition accuracy on the TIMIT SX sentences ranging from 73.0% to 75.5%, which is slightly higher than the proposed model. This may be partly due to the fact that Schuster and Paliwal, like in most other studies, reduced the original 61 phoneme labels to a compact set of 45 labels, which in turn reduced the errors of some highly confusable phonemes, e.g., nasals. In the literature, phoneme recognition accuracy as high as 75.6% has been reported on the TIMIT database using the tandem MLPs with an adjustable neural structure (Schwarz et al., 2006). The number of neurons in the hidden layer of neural networks was increased until the PER saturated (Schwarz et al., 2006).

Table 5.6 summarizes the phoneme recognition accuracy of different ASR systems on the TIMIT SX sentences, where the 39 dimensional MFCCs are used at the front end and the accuracy of 45 phonemes are measured in the results. The phoneme recognition accuracy for the proposed API recognizer is 74.4%, shown in the last row of Table 5.6. The observation is that the performance of ASR can be improved in several ways.

1. Increasing the number of Gaussian mixture (GM) in the HMM baseline: 40 GM (74.5%) vs. 16 GM (59.2%), (method 1 and 2). This is the conventional method to improve ASR performance.

2. Using NNs to replace the HMM (method 3). This is a paradigm shift, where the best reported in the literature is the tandem approach: 75.6%, slightly higher than the best HMM. A more interesting approach is using RNNs to model the temporal dynamics of speech features implicitly instead of using the FFNNs (or MLPs). The bid-directional RNN (method 4) obtains phone accuracy of 73.0%, which uses both future and past recurrent links/memories to optimize the capability of the paradigm. The performance is reasonable, yet the input features are purely acoustic: MFCCs, same as the HMM.

3. Using a hybrid HMM/NN approach (method 5), where NNs interpolate the phone posterior of the HMM: HMM+MLP (64.8%). The system accuracy is not as high as the 40 GM, but it improves the baseline 16 GM by 5.3% absolute under the exact same testing condition, shown in row 5 of Table 5.

4. Using additional knowledge sources to improve the features representation (method 5, 6, and 7). However, the pseudo-articulatory features (PAFs) have shown promising results, when used in the hybrid HMM/NN and the RNN framework, (method 5 and 6). Yet the improvement is usually limited. In the literatures, the PAFs are first transcribed by the phonological rules such as the manner and place of articulation of phones, then they are used to rescore the HMM baseline or to calculate the phone posterior. In other words, the amount of production knowledge is proportional to the phonological codebook. They seem more reliable than the articulatory codebook for the APFs. However, they may simply reduce to a redistribution of the acoustic features during ASR.

5. The APFs in this study (method 7) are estimated from the acoustic signal. In particular, the synthetic data after clustering is used for the acoustic-to-articulatory mapping. The APFs offer a unified explanation using the two aspects of the human speech: perception (BFCCs) and production (RATs).

There are mainly two types of contributions for ASR improvement in the above methods: one is the different paradigms/algorithms (HMMs vs. NNs), the other

is the use of speech production knowledge (MFCCs vs. PAFs and APFs). Table 5.6 shows that the production knowledge is an important contributor to the speech recognition performance gain, and the neural topology also plays a complementary role to the proposed articulatory features. However, the phoneme accuracy shows a bottleneck below 80%. This is not satisfactory compared with human recognition performance, where current ASR systems usually apply syntax and semantic rules to embed linguistic constraints for improved phoneme accuracy. These will also interesting for future studies.

Table 5.6: Summary of phoneme recognition accuracy (%) on the TIMIT sentences in the literature.

| Structure | | Dataset | Phoneme accuracy |
|---|---|---|---|
| 1. HMM (40 GM per state) | (Jeon and Juang, 2007) | SX, SI | 74.5 |
| 2. HMM (16 GM per state) | (Siniscalchi and Lee, 2009) | SX, SI | 59.5 |
| 3. Tandem MLP | (Schwarz et al., 2006) | SX, SI | **75.6** |
| 4. RNN | (Schuster and Paliwal, 1997) | SX | 73.0 |
| 5. HMM+MLP (PAF) | (Siniscalchi and Lee, 2009) | SX, SI | 64.8 |
| 6. RNN (PAF) | (King and Taylor, 2000) | SX, SI | 63.3 |
| 7. API (APF) | (Huang and Er, 2011) | SX | 74.4 |

### 5.6.2 Phoneme Error Patterns

The proposed neural model exhibits different phoneme error patterns. Table 5.7 reports the top and the bottom phoneme errors obtained by the HMM baseline and the API model during the clean testing with $SNR = 30dB$. The acoustic baseline has lower accuracy for the nasals and the fricatives than the articulatory model. These phonemes have noise-like qualities which are difficult to distinguish using the acoustic features (Scharenborg et al., 2007). There are fewer frame labels for nasals: /eng, em/, fricatives: /f, z, s, sh/, and plosive closures: /dcl, tcl/, compared to other phonemes, which introduces data sparsity during training. In the HMM baseline, these phoneme results in either the highest or the lowest error. In other words, their accuracy are highly variant in the acoustic baseline.

The issue is less severe for the API model. Only few of the sparsely distributed phonemes /eng, s, sh/ appear in the top or bottom errors. There are two reasons. First, the 45 broad phoneme labels and the 255 narrow ones are mapped with the 13 APFs, so there are at least three times more frames in the training data for the articulatory model. Second, the articulatory dynamics in the APFs contribute to alter the phoneme error patterns. Previously the combined use of acoustic and articulatory features have shown to utilize the two knowledge sources and improve ASR performance (Huang and Er, 2011). However, the APFs are not without restrictions, as evidenced by the high error rates of the liquids: /l, w/ and the glottal sound: /hh/. The main issue is that the synthetic database generated by the 2-D bio-mechanical system has difficulty in describing the complex tongue and glottal movements. These phonemes require additional physiological parameters, for example, in a 3-D or 2-D vocal tract with fricative excitation sources at the specific tongue locations (Birkholz et al., 2007).

Table 5.7: Phoneme recognition accuracy (%) obtained by the HMM baseline and the API model.

| | HMM | | API | |
|---|---|---|---|---|
| Top 5 errors | eng | 5.1 | uh | 19.0 |
| | uh | 21.0 | eng | 21.6 |
| | em | 27.0 | l | 26.9 |
| | ih | 33.5 | w | 31.0 |
| | dcl | 39.0 | hh | 32.5 |
| Bottom 5 errors | ao | 91.0 | aa | 89.8 |
| | f | 92.4 | ao | 92.7 |
| | z | 93.2 | ch | 93.3 |
| | sh | 95.2 | s | 98.0 |
| | s | 97.7 | sh | 98.5 |

Fig. 5.7 shows the phonetic analysis of an utterance extracted from the test set, the consonant-vowel pattern “by” in the sentence “Jane may earn more money by working hard” (SX4) by a male speaker (MAE). The upper panel shows the recorded acoustic waveform. The middle panel shows the linear frequency (0 to

3000 Hz) spectrogram after FFT before the non-linear mel or Bark scale warping. The lower panel shows the broad and narrow phonetic annotations. In the narrow annotation, the voiced plosive /b/ has three stages of articulation during the consonantal onset: lip flapping [bv], vocalic silence [bc], and voicing [ba] followed by the vowel sound. Similarly the diphthong /ay/ has two stages of articulation, which are identified by the disappearance of the third formant from the /a/ to the /y/ segments, shown in the narrow annotation as [aIa] and [aII]. Compared with the broad annotation as used in the acoustic HMM baseline, the API model embed more descriptive phonetic details in the narrow annotations. The details enable the API model to infer the underlying articulatory dynamics for accurate phoneme classification.

Fig. 5.8 shows three estimated APFs: JH, $LL_y, TB_y$, which are normalized to the [0, 1] interval. The dashed lines represent the articulatory commands in the synthetic training data, and the solid line represents the estimated APFs in the API model. Abrupt articulatory movements are observed at cross sections of the phonetic intervals. The step like shape resembles the binary PAFs, where the presence/absence of a feature is denoted as 1/0 or +/− (King et al., 2007; Scharenborg et al., 2007). However, for the continuous APFs there also exist intermediate values in Fig. 5.8, which give relative rather than absolute measures as used in the discrete PAFs. The measure better describes the phonetic variations, since phonemes are usually articulated without reaching the final targets, especially in continuous speech (Fang et al., 2009).

The API model calculates the probabilities of the output phonemes based on the estimated APFs which are conditioned on the input acoustic features. Besides the annotation details in the phonetic layer, BFCCs in the input layer are competent in preserving the acoustic information in clean and noisy conditions. Fig. 5.9 shows the mel and the Bark cepstrum. Consonant-vowel pattens are the building blocks of English words and sentences. The coarticulation effects at the consonant onset and at the vowel formants transition are visible in the two cepstrums. In clean condition, the two types of acoustic features, MFCCs and BFCCs, are closely matched along the time and the frequency-axes. Previous experiments have shown that their performance were equivalent on the same HMM recognizer (Huang and Er, 2012c). In noisy condition, white Gaussian

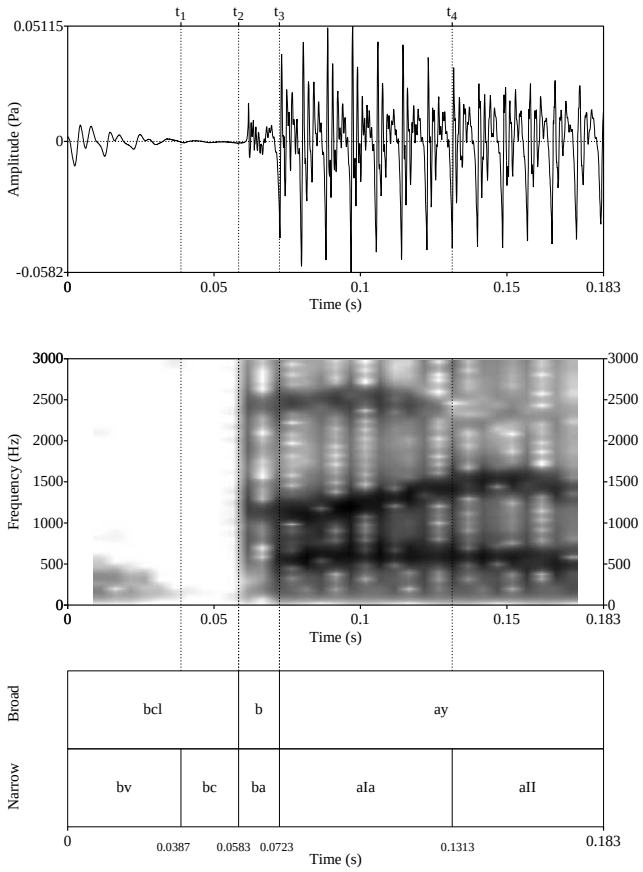



Figure 5.7: The phonetic analysis of the consonant-vowel pattern /bay/ for the word"by" in the natural speech corpus, the acoustic waveform (upper panel), the linear frequency spectrogram (middle panel), and the broad/narrow phonetic annotations (lower panel). The dotted vertical lines mark the phonetic boundaries at $t_1 = 0.0387$ sec, $t_2 = 0.0583$ sec, $t_3 = 0.0723$ sec, and $t_4 = 0.1313$ sec.

with $SNR = 15$ dB, the Bark filters give slightly better amplitude compression than the mel filters in the plosive closure interval $(t_1 - t_2)$, as seen by the white squares in Fig. 5.9.

## 5.7 Summary

This chapter describes the neural based API model, which retrieves and utilizes the production knowledge based APFs to improve speech recognition per-

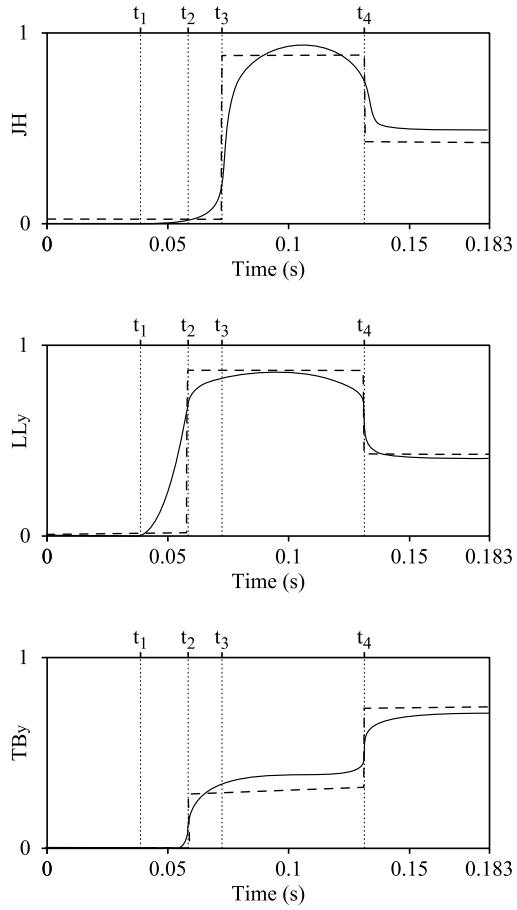



Figure 5.8: Three estimated APFs: JH, $LL_y$, and $TB_y$ in the API model (solid line). The dashed lines represent the articulatory configurations in the synthetic corpus.

formance. It also generates the synthetic corpus with the articulatory-acoustic mapping information using the biomechanical speech synthesizer. The performance of the inversion module is evaluated on the synthetic and the natural speech data. Initial results indicate that the inversion module obtains accurate articulatory estimates. The phoneme recognition performance demonstrates that the API model are more competent in modeling highly variant phonetic events than the HMM baseline. It outperforms the acoustic HMM baseline with improved speaker independence and noise robustness. Furthermore, the API system is tuned with the same set of speech data in an off-line learning mode before

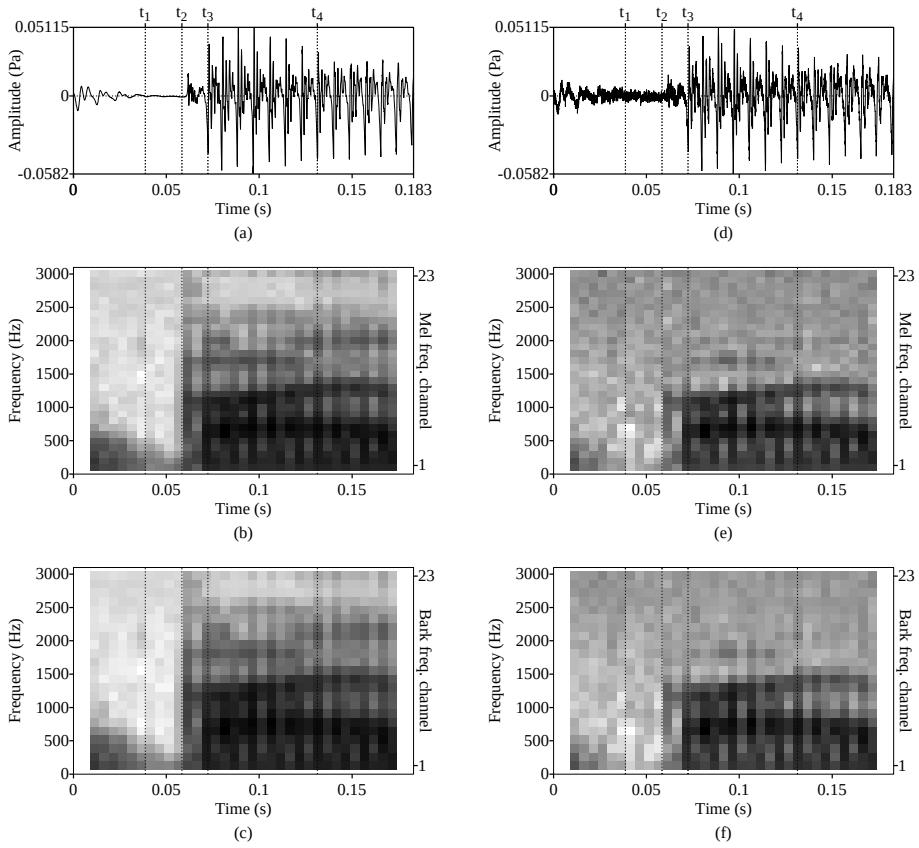



Figure 5.9: Comparison of the mel and the Bark-scale cepstral measures on the /bay/ sound: (b) & (c) in clean and (e) & (f) in noisy condition, added white Gaussian with $SNR = 15dB$, respectively.

on-line testing. Thus portability and computational efficiency are another two salient properties of the proposed neural model.

# Chapter 6

# Conclusion

Speech processing technology continues to fascinate engineers and researchers in human-machine interaction. It leads to many promising grounds as well as challenging tasks (Juang and Furui, 2000). The use of production knowledge in speech recognition is one such task. This work describes a method with three stages to embed the articulatory phonetic features for improved speech recognition. First the neural controller in Chapter 3 tracks the articulatory movements of the human vocal tract and infers the activation patterns of the underlying muscular structures. It is able to manipulate the mass-spring based elastic tract walls in a 2-D articulatory synthesizer to realize speech motor control and to reproduce the articulatory-acoustic mapping of English phonemes. It achieves high accuracy during on-line tracking of the vocal apparatus in the simulation of consonant-vowel sequences. Next the non-uniform segmentation method is used in Chapter 4 to build the English pronunciation models. The broad phoneme forms and the narrow phonetic forms are used to annotate the variations in the acoustic signal. Experimental results show that the articulatory feature space presents a much smaller variance than the acoustic feature space. Finally the neural base articulatory phonetic inversion (API) model is implemented in Chapter 5. The model retrieves the articulatory phonetic features (APFs) from the acoustic features. It achieves improved recognition accuracy and robustness in two diverse conditions: with different speakers and in noisy environments.

Compared with the existing phonological articulatory features (PAFs) which are derived from the broad linguistic definitions, e.g., manner of articulation

(MOA) and place of articulation (POA), as used in (Frankel et al., 2007; Kirchhoff et al., 2002; Siniscalchi and Lee, 2009), the proposed APFs use a more reliable heuristic mapping strategy to retrieve the pronunciation details on a set of hand-labeled sentences. The proposed API approach roots in the concept that the speech sound occupies a spread region, rather than isolated points, in the auditory and the articulatory domain (Damper and Harnad, 2000; Kielar et al., 2011; Mottonen and Watkins, 2009). What differs this study from others is the *unified explanation* of speech events in the production and the perception domains. The proposed neural module distinguishes the phonetic base-forms from the surface-forms, i.e., the pronunciation variations of English speech. It use two knowledge sources that are not present in conventional classifiers, which are analogous human speech processing, i.e., the listener-oriented maximization of auditory discriminations and the speaker-oriented minimization of articulatory effort.

In addition, the research work addresses the non-uniqueness and the non-linearity issue in the inversion experiments by incorporating the production knowledge at three places. First, in the control model (cf. Chapter 3), the biomechanical synthesizer approximates the human anatomy in physiological and functional properties of speech production. Second, in the pronunciation model (cf. Chapter 4), the heuristic learning algorithm mimics the experience of human speech acquisition. Third, in the inversion model (cf. Chapter 5), the data clustering algorithm minimizes the within-class scatter distance and maximizes the across-class scatter distance in the synthetic data, which is also analogous to the categorical nature of human speech perception.

## 6.1 Recommendation for Further Research

Much remains to be done in speech recognition as well as in articulatory synthesis. One near goal is to design a fully functional articulatory speech synthesizer, as shown in Fig. 6.1. At present there are few synthesis systems which accommodate both the articulatory and the acoustic modules, even fewer which include a control module. The adaptive control model presented in this thesis is a first step toward

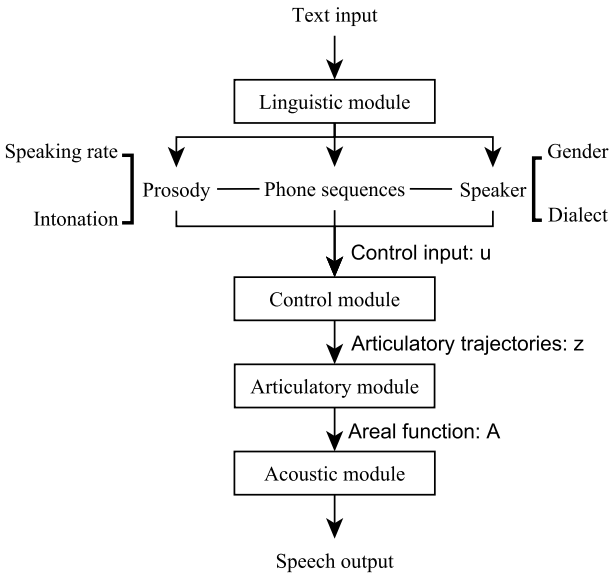



Figure 6.1: Information flow in a fully functional TTS system with the adaptive controller.

an automatically controlled and self-tuned articulatory synthesizer. When using the adaptive controller for TTS synthesis, the phone sequence, the speaking rate, and the prosodic information need to be specified by the adaptive controller to generate the corresponding articulatory trajectories. It can also include an auditory feedback mechanism to perceive the above acoustic signals like a human listener.

Another future goal is to design an improved speech recognizer that not only generates the best word string but also outputs a set of multiple hypotheses that can be represented as a word lattice or an n-best list. It is often observed that human speech recognition is more dependent upon linguistic knowledge, context information and other post-recognition processing than previously supposed (De Mori et al., 2008). This may as well explain the superiority of human speech recognition compared to machine based systems. When lattices of word hypotheses are generated, it is likely that the uttered words are somewhere in the lattice, making it possible to obtain coherent semantic hypotheses from these hypothe-

ses. Fig. 6.2 shows one such system. Using semantic parsing, word hypotheses can be dynamically attached to non-terminal symbols For example, Fig. 6.3 demonstrates a ticket query system using the speech recognizer and a semantic parser. The semantic parsing tree in Fig. 6.3 uses node and hierarchical links with extension scores, which are based on the assumption that sentences that are grammatically correct. For context-dependent applications, e.g. the flight inquiry, it is important to use confidence measures that integrate information related to the whole dialog context rather than just the acoustic signal. These will be interesting for future research.

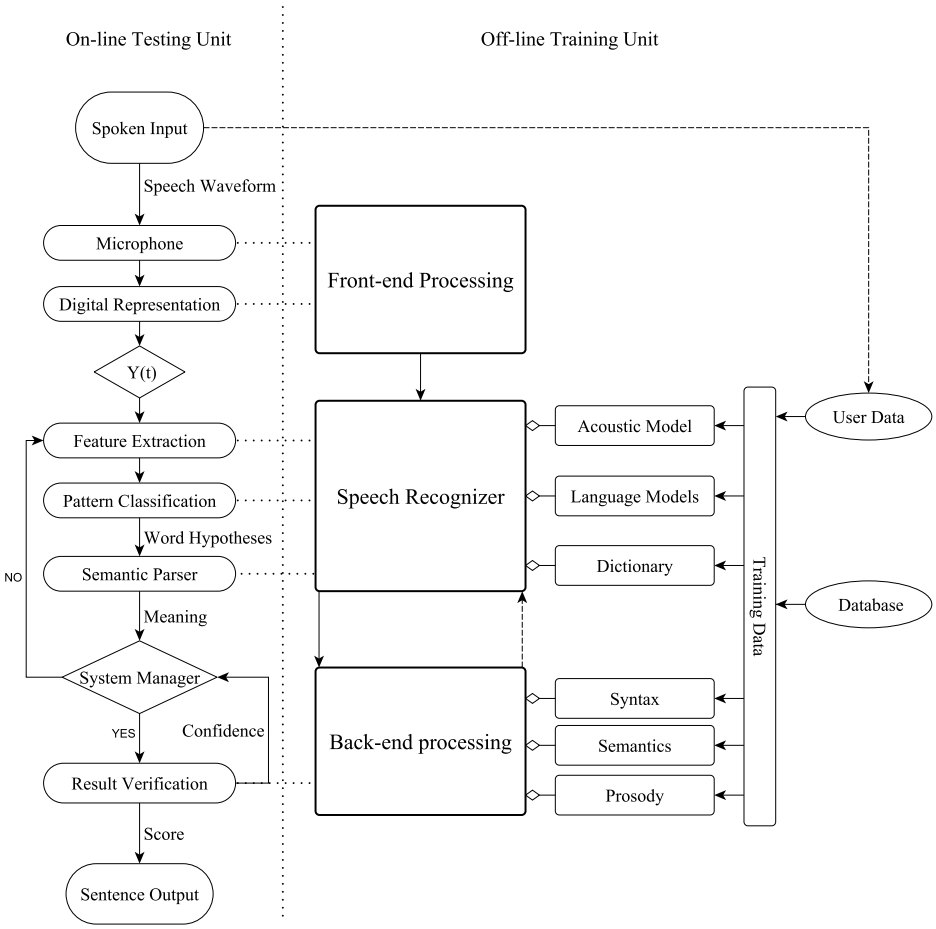



Figure 6.2: Structure of the improve ASR system.

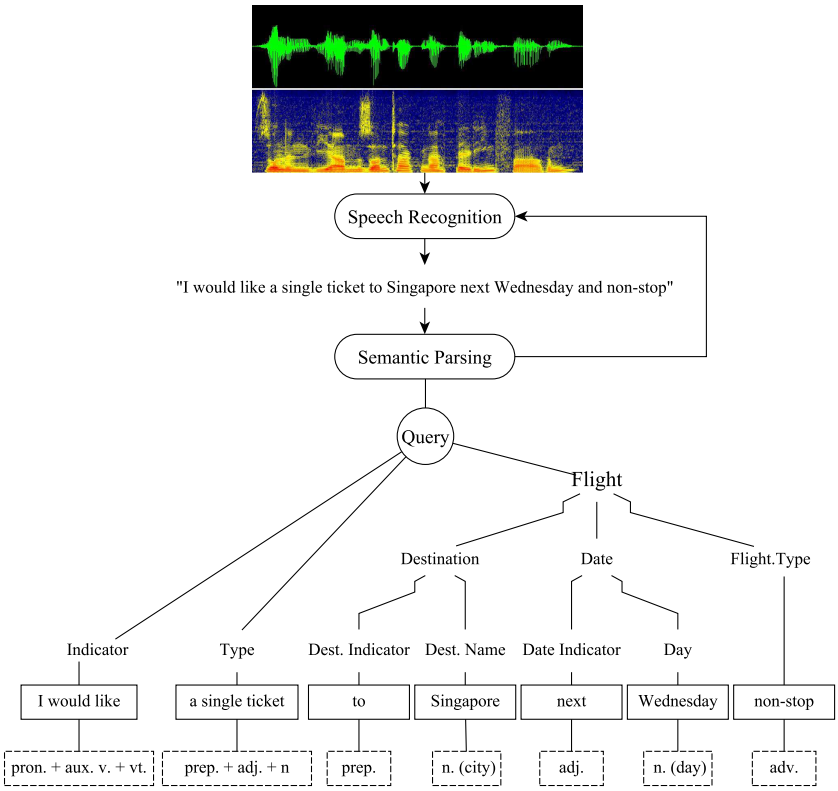



Figure 6.3: Semantic parsing in an information retrieval system.

# Author's Publications

## Journal Articles

1. Huang Guangpu & Er Meng Joo, An Adaptive Neural Control Scheme for Articulatory Synthesis of Consonant-Vowel Sequences, Computer Speech and Language. Impact factor: 1.319. To appear.

2. Huang Guangpu & Er Meng Joo, Model-Based Articulatory Phonetic Inversion for Improved Speech Recognition, Speech Communication. Impact factor: 1.267. To appear.

3. Huang Guangpu & Er Meng Joo, Articulatory Based Multi-Dimensional Pronunciation Modeling for Robust Speech Recognition, submitted to submitted to IEEE Transactions on Audio Speech and Language Processing. Impact factor: 1.498.

## Conference Articles

1. Huang Guangpu & Er Meng Joo, An Adaptive Control Scheme for Articulatory Synthesis of Plosive-Vowel Sequences, in Proceedings of the 38th Annual Conference of the IEEE Industrial Electronics Society (IECON 2012), pages 1465-1470.

2. Huang Guangpu & Er Meng Joo, Model-based Articulatory Phonetic Features for Improved Speech Recognition, in Proceedings of the 2012 IEEE

World Congress on Computational Intelligence (IJCNN 2012), Brisbane, Australia, pages 1-8.

3. Huang Guangpu & Er Meng Joo, Bi-directional Phonetic Modeling of Consonant-Vowel Speech Patterns, in Proceedings of the 7th IEEE Conference on Industrial Electronics and Applications (ICIEA 2012), Singapore, pages 1798-1803.

4. Huang Guangpu & Er Meng Joo, Combined Articulatory and Auditory Processing for Improved Speech Recognition, in Proceedings of the 7th IEEE Conference on Industrial Electronics and Applications (ICIEA 2012), Singapore, pages 972-977.

5. Huang Guangpu & Er Meng Joo, A Novel Neural-Based Pronunciation Modeling Method for Robust Speech Recognition, IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011), Big island, Hawaii, USA, pages 517-522.

6. Huang Guangpu & Er Meng Joo, Robust Phoneme Recognition Based on Auditory Processing and Hidden Markov Modeling, accepted for presentation by the 8th International Symposium on Neural Networks (ISNN 2011), Guilin, China. 2011.

7. Huang Guangpu & Er Meng Joo, A Hybrid Computational Model for Spoken Language Understanding, in Proceedings of the 11th International Conference on Control, Automation, Robotics and Vision (ICARCV 2010), Singapore, pages 2078-2083.

8. Huang Guangpu & Er Meng Joo, Design and Development of a Student Assistance System for Speaking Standard English, accepted for presentation by the 4th IEEE Conference on Industrial Electronics and Applications (ICIEA 2009), Xi'an, China. 2009.

# Bibliography

Afify, M., Cui, X., and Gao, Y. (2009). Stereo-based stochastic mapping for robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1325 –1334.

Ali, A., Bhatti, S., and Mian, M. S. (2006). Formants based analysis for speech recognition. In *Engineering of Intelligent Systems, 2006 IEEE International Conference*.

Allen, J. (1996). Harvey fletcher's role in the creation of communication acoustics. *Journal of the Acoustical Society of America*, 99(4 I):1825–1839.

Allen, J. (2008). Nonlinear cochlear signal processing and masking in speech perception. *Springer Handbook on Speech Processing and Speech Communication*, pages 27–60.

Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2):637–655.

Badin, P., Bailly, G., Reveret, L., Baciu, M., Segebarth, C., and Savariaux, C. (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on mri and video images. *Journal of Phonetics*, 30(3):533–553.

Baer, T., Alfronso, P. J., and Honda, K. (1988). Electromyography of the tongue muscle during vowels in /epvp/ environment. *Annual Bulletin of the Research Institute of Logopaedics and Phoniatrics*, 22:7–19.

Bailly, G. (1997). Learning to speak. sensori-motor control of speech movements. *Speech Communication*, 22(2-3):251–267.

Bailly, G., Laboissiere, R., and Galvain, A. (1997). Learning to speak: Speech production and sensori-motor representations. *Advances in Psychology*, 119(C):593–615.

Behbood, H., Seyyedsalehi, S., Tohidypour, H., Najafi, M., and Gharibzadeh, S. (2011). A novel neural-based model for articulatory-acoustic inversion mapping. *Neural Computing and Applications*, pages 1–9.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10-11):763–786.

Bickford, A. (2006). *Articulatory Phonetics: Tools For Analyzing The World's Languages*. Summer Institute of Linguistics, 4th edition.

Birkholz, P. (2005). *3-D Articulatory Speech Synthesis*. PhD thesis, University of Rostock, Rostock, Germany.

Birkholz, P. and Jackel, D. (2004). Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. In *INTERSPEECH*. ISCA.

Birkholz, P., Jackel, D., and Kröger, B. J. (2007). Simulation of losses due to turbulence in the time-varying vocal system. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1218–1226.

Birkholz, P., Kröger, B., and Neuschaefer Rube, C. (2011). Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Transactions on Audio, Speech and Language Processing*, 19(5):1422–1433.

Blumstein, S., Stevens, K., and Nigro, G. (1977). Property detectors for bursts and transitions in speech perception. *Journal of the Acoustical Society of America*, 61(5):1301–1313.

Boersma, P. (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. PhD thesis, University of Amsterdam.

Boets, B., Wouters, J., van Wieringen, A., and Ghesquiere, P. (2007). Auditory processing, speech perception and phonological ability in pre-school children at high-risk for dyslexia: A longitudinal study of the auditory temporal processing theory. *Neuropsychologia*, 45(8):1608–1620.

Bortman, M. and Aladjem, M. (2009). A growing and pruning method for radial basis function networks. *IEEE Transactions on Neural Networks*, 20(6):1039–1045.

Bouabana, S. and Maeda, S. (1998). Multi-pulse lpc modeling of articulatory movements. *Speech Communication*, 24(3):227–248.

Bourlard, H. and Morgan, N. (1993). *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic.

Browman, C. P. and Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49:155–180.

Brown, G. and Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech and Language*, 8:297–336.

Buchaillard, S., Perrier, P., and Payan, Y. (2009). A biomechanical model of cardinal vowel production: Muscle activations and the impact of gravity on tongue positioning. *Journal of the Acoustical Society of America*, 126(4):2033–2051.

Che, C., Lin, J., Pearson, J., de Vries, B., and Flanagan, J. (1994). Microphones arrays and neural networks for robust speech recognition. In *ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Chomsky, N. (2006). *Language and Mind*. Cambridge University Press.

Claes, T., Dologlou, I., ten Bosch, L., and van Compernolle, D. (1998). A novel feature transformation for vocal tract length normalization in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 6(6):549 – 557.

Cole, R., Mariani, J., Uszkoreit, H., Varile, G. B., Zaenen, A., Zampolli, A., and Zue, V. (1997). Survey of the state of the art in human language technology.

Cook, P. (1990). Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing. Master's thesis, Stanford University, Stanford, California.

Cox, R. V. (2000). Speech and language processing for next-millenium communication services. In *Proceedings of the IEEE*, volume 88.

Cui, X. and Alwan, A. (2005). Noise robust speech recognition using feature compensation based on polynomial regression of utterance snr. *IEEE Transactions on Speech and Audio Processing*, 13(6):1161–1172.

Damper, R. and Harnad, S. (2000). Neural network modeling of categorical perception. *Perception and Psychophysics*, 62 (4):843–867.

Davis, M. and Johnsrude, I. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, 229(1-2):132–147.

De Mori, R., Bechet, F., Hakkani-Tur, D., McTear, M., Riccardi, G., and Tur, G. (2008). Spoken language understanding. *Signal Processing Magazine, IEEE*, 25(3):50 –58.

Dede, G. and SazlI, M. H. (2009). Speech recognition with artificial neural networks. *Digital Signal Processing*, In Press, Corrected Proof.

Deng, L., Droppo, J., and Acero, A. (2004). Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features. *IEEE Transactions on Speech and Audio Processing*, 12(3):218–233.

Duck, F. A. (1990). *Physical Properties of Tissues: A Comprehensive Reference Book*. Academic Press, London.

Elman, J. (1988). Finding structure in time. Technical report, Center for Research in Language, University of California, San Diego.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Er, M. J., Chen, W., and Wu, S. (2005). High-speed face recognition based on discrete cosine transform and RBF neural networks. *IEEE Transactions on Neural Networks*, 16(3):679–691.

Er, M. J. and Gao, Y. (2003). Robust adaptive control of robot manipulators using generalized fuzzy neural networks. *IEEE Transactions on Industrial Electronics*, 50(3):620–628.

Fang, Q. (2009). *A Study On Construction And Control Of A Three-Dimensional Physiological Articulatory Model For Speech Production*. PhD thesis, School of Information Science, Japan Advanced Institute of Science and Technology.

Fang, Q., Fujita, S., Lu, X., and Dang, J. (2009). A model-based investigation of activations of the tongue muscles in vowel production. *Acoustical Science and Technology*, 30(4):277–287.

Feldman, A. G. (1986). Once more on the equilibrium-point hypothesis (lambda model) for motor control. *Journal of motor behavior*, 18(1):17–54.

Flanagan, J., Ishizaka, K., and Shipley, K. (1975). Synthesis of speech from a dynamic model of the vocal cords and vocal tract. *The BELL System Technical Journal*, 54(3):485–506.

Flanagan, J. L. (1972). *Speech analysis, synthesis, and perception*. Berlin: Springer-Verlag.

Flynn, R. and Jones, E. (2008). Combined speech enhancement and auditory modelling for robust distributed speech recognition. *Speech Communication*, 50(10):797–809.

Frankel, J., Wester, M., and King, S. (2007). Articulatory feature recognition using dynamic bayesian networks. *Computer Speech and Language*, 21(4):620–640.

Fry, D. B. (1959). Theoretical aspects of the mechanical speech recognition. *Journal of British Institution of Radio Engineers*, 19(4):pp. 211–229.

Gajic, B. and Paliwal, K. (2006). Robust speech recognition in noisy environments based on subband spectral centroid histograms. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):600–608.

Gao, Y. and Er, M. J. (2003). Online adaptive fuzzy neural identification and control of a class of MIMO nonlinear systems. *IEEE Transactions on Fuzzy Systems*, 11(4):462–477.

Gao, Y. and Er, M. J. (2005). An intelligent adaptive control scheme for post-surgical blood pressure regulation. *IEEE Transactions on Neural Networks*, 16:475–483.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia, USA.

Gemmeke, J., Van Hamme, H., Cranen, B., and Boves, L. (2010). Compressive sensing for missing data imputation in noise robust speech recognition. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):272 –287.

Goddeau, D. and Zue, V. (1992). Integrating probabilistic lr parsing into speech understanding systems. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1.

Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech Communication*, 16(3):261–291.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96:280–301.

Halliday, M. A. K. and Webster, J. (2006). *On Language and Linguistics*. Continuum International Publishing Group.

Heinz, J. and Stevens, K. (1961). On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America*, 33(5):589–596.

Hermansky, H. and Morgan, N. (1994). Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578 –589.

Hieronymus, J., Alexander, M., Bennett, C., Cohen, I., Davies, D., Dalby, D., Laver, J., Barry, W., Fourcin, A., and Wells, J. (1990). Proposed speech segmentation criteria for the SCRIBE project. Technical report, University College London.

Hirayama, M., Vatikiotis-Bateson, E., and Kawato, M. (1993). Inverse dynamics of speech motor control. In *NIPS*, pages 1043–1050.

Hixon, T. J. (1987). *Respiratory function in speech and song*. London: Taylor & Francis.

Holmberg, M., Gelbart, D., and Hemmert, W. (2006). Automatic speech recognition with an adaptation model motivated by auditory processing. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):43–49.

Huang, G. and Er, M. J. (2010). A hybrid computational model for spoken language understanding. In *Proceedings of the 11th International Conference on Control Automation Robotics Vision (ICARV)*, pages 2078–2083.

Huang, G. and Er, M. J. (2011). A novel neural-based pronunciation modeling method for robust speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Hawaii, USA*, pages 517–522.

Huang, G. and Er, M. J. (2012a). An adaptive control scheme for articulatory synthesis of plosive-vowel sequences. In *Proceedings of the 38th Annual Conference of the IEEE Industrial Electronics Society (IECON)*, pages 1465–1470.

Huang, G. and Er, M. J. (2012b). Bi-directional phonetic modeling of consonant-vowel speech patterns. In *Proceedings of the 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 1798–1803.

Huang, G. and Er, M. J. (2012c). Combined articulatory and auditory processing for improved speech recognition. In *Proceedings of the 7th IEEE Conference on Industrial Electronics and Applications (ICIEA'12)*, pages 972–977.

Huang, G. and Er, M. J. (2012d). Model-based articulatory phonetic features for improved speech recognition. In *Proceedings of the 2012 IEEE World Congress on Computational Intelligence (WCCI)*, pages 1–8.

Huang, K.-C. and Juang, Y.-T. (2003). Feature weighting in noisy speech recognition. *Electronics Letters*, 39(12):938–939.

Huckvale, M. (2004). Scribe-spoken corpus recordings in british english. online: http://www.phon.ucl.ac.uk/resource/scribe/scribe-manual.htm.

Hughes, G. and Halle, M. (1956). Spectral properties of fricative consonants. *Journal of the Acoustical Society of America*, 28(2):303–310.

Ishizaka, K., French, J. C., and Flanagan, J. L. (1975). Direct determination of vocal tract wall impedance. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23:370–373.

Jankowski, Jr., C., Vo, H.-D., and Lippmann, R. (1995). A comparison of signal processing front ends for automatic word recognition. *IEEE Transactions on Speech and Audio Processing*, 3(4):286 –293.

Jelinek, F. (1990). Self-organized language modeling for speech recognition. *Readings in Speech Recognition*, pages 450–506.

Jeon, W. and Juang, B.-H. (2007). Speech analysis in a model of the central auditory system. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(6):1802–1817.

Jones, D. (1972). *An outline of English phonetics*. Cambridge: W. Heffer & Sons Ltd.

Jordan, M. I. (1986). Serial order in behavior: a parallel distributed processing approach. Technical Report 8604, San Diego: University of California, Institute for Cognitive Science.

Juang, B. H. and Furui, S. (2000). Automatic recognition and understanding of spoken language-a first step toward natural human-machine communication. *Proceedings of the IEEE*, 88(8):1142–1165.

Kanazawa, H., Tachimori, M., and Takebayashi, Y. (1995). A hybrid wordspotting method for spontaneous speech understanding using word-based pattern matching and phoneme-based hmm. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 289–292 vol.1.

Kawato, M., Maeda, Y., Uno, Y., and Suzuki, R. (1990). Trajectory formation of arm movement by cascade neural network model based on minimum torque-change criterion. *Biological Cybernetics*, 62:275–288.

Kelly, J. L. and Lochbaum, C. (1962). Speech synthesis. In *Proceedings of the Speech Communications Seminar, paper F7. Stockholm, Speech Transmission Laboratory, Royal Institute of Technology.*

Kelso, J. A. S., Saltzman, E. L., and Tuller, B. (1986). The dynamical perspective on speech production: data and theory. *Journal of Phonetics*, 14:29–59.

Kielar, A., Milman, L., Bonakdarpour, B., and Thompson, C. (2011). Neural correlates of covert and overt production of tense and agreement morphology: Evidence from fmri. *Journal of Neurolinguistics*, 24(2):183–201.

Kim, N. S., Kim, Y. J., and Kim, H. W. (2004). Feature compensation based on soft decision. *Signal Processing Letters, IEEE*, 11(3):378–381.

King, R. (1997). New challenges in automatic speech recognition and speech understanding. In *TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications., Proceedings of IEEE*, volume 1, page 287 vol.1.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., and Wester, M. (2007). Speech production knowledge in automatic speech recognition. *Journal of the Acoustical Society of America*, 121(2):723–742.

King, S. and Taylor, P. (2000). Detection of phonological features in continuous speech using neural networks. *Computer Speech and Language*, 14(4):333–353.

Kinjo, T. and Funaki, K. (2006). On hmm speech recognition based on complex speech analysis. In *32nd Annual Conference on IEEE Industrial Electronics, IECON 2006*, pages 3477–3480.

Kirchhoff, K., Fink, G. A., and Sagerer, G. (2002). Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, 37(3-4):303–319.

Klatt, D. (1977). Review of the arpa speech understanding project. *Journal of the Acoustical Society of America*, 62(6):1345–1366.

Kröger, B., Graf-Borttscheller, V., and Lowit, A. (2008). Two and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders. In *Interspeech, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia.*

Kröger, B., Kannampuzha, J., and Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Communication*, 51(9):793 –809.

Kröger, B., Schroder, G., and Opgen-Rhein, C. (1995). A gesture-based dynamic model describing articulatory movement data. *Journal of the Acoustical Society of America*, 98(4):1878–1889.

Kuhn, R. and De Mori, R. (1995). The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):449–460.

Lang, K. J., Waibel, A. H., and Hinton, G. E. (1990). A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3(1):23–43.

Lee, C. H., Giachin, E., Rabiner, L. R., Pieraccini, R., and Rosenberg, A. E. (1992). Improved acoustic modeling for large vocabulary continuous speech recognition. *Computer Speech & Language*, 6(2):103–127.

Lee, C. H., Rabiner, L. R., Pieraccini, R., and Wilpon, J. G. (1990). Acoustic modeling for large vocabulary speech recognition. *Computer Speech and Language*, 4(2):127–165.

Lee, Y. C., Chen, H. H., and Sun, G. Z. (1988). A neural network approach to speech recognition. *Neural Networks*, 1(Supplement 1):306–306.

Lesser, V. R., Fennell, R. D., Erman, L. D., and Reddy, D. R. (June 1975). Organization of the hearsayii: Speech understanding system. *IEEE Transactions on Acoust., Speech, Signal Processing*, ASSP-23:1123.

Levelt, W. J. (1999). Models of word production. *Trends in cognitive sciences*, 3(6):223–232.

Li, F., Menon, A., and Allen, J. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *Journal of the Acoustical Society of America*, 127(4):2599–2610.

Liberman, A. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America*, 29(1):117–123.

Lin, C.-M. and Li, H.-Y. (2012). Tsk fuzzy cmac-based robust adaptive backstepping control for uncertain nonlinear systems. *IEEE Transactions on Fuzzy Systems*, page in press.

Lippmann, R. (1987). An introduction to computing with neural nets. *ASSP Magazine, IEEE*, 4(2):4–22.

Lippmann, R. P. (1989). Review of neural networks for speech recognition. *Neural Comput.*, 1(1):1–38.

Löfqvist, A. and Gracco, V. (2002). Control of oral closure in lingual stop consonant production. *Journal of the Acoustical Society of America*, 111(6):2811–2827.

Lubensky, D. M., Asadi, A. O., and Naik, J. M. (1994). Connected digit recognition using connectionist probability estimators and mixture-gaussian densities. In *In Proceedings of the 1994 International Conference on Spoken Language Processing, Yokohama, Japan.*, volume 1, pages 295–298.

Maeda, S. (1982). A digital simulation method of the vocal-tract system. *Speech Communication*, 1(3, A14):199–229.

Malecot, A. (1973). Computer-assisted phonetic analysis techniques for large recorded corpuses of natural speech. *Journal of the Acoustical Society of America*, 53:356.

Martin, J. H. and Jurasfsky, D. (2008). *Speech and language processing*. Prentice Hall.

Merhav, N. and Lee, C.-H. (1993). A minimax classification approach with application to robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(1):90 –100.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53:1070–1082.

Millar, J., Vonwiller, J., Harrington, J., and Dermody, P. (1994). The australian national database of spoken language. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94.*, volume i, pages I 97–I100 vol.1.

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., and Goldstein, L. (2011). Articulatory information for noise robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1913–1924.

Mottonen, R. and Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *Journal of Neuroscience*, 29:9819–9825.

Muller, D., de Siqueira, M., and Navaux, P. (2006). A connectionist approach to speech understanding. In *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, pages 3790–3797.

Nasibov, Z. and Kinnunen, T. (2012). Decision fusion of voice activity detectors. Master's thesis.

Nelson, W. L. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics*, 46:135–147.

Niyogi, P. and Ramesh, P. (2003). The voicing feature for stop consonants: Recognition experiments with continuously spoken alphabets. *Speech Communication*, 41(2-3):349–367.

Nunn, J. (1993). *Applied Respiratory Physiology*. Butterworth-Heinemann.

Ogata, K. and Sonoda, Y. (2003). Reproduction of articulatory behavior based on the parameterization of articulatory movements. *Acoustic Science and Technology*, 24:403–405.

Parham, A., Guangji, S., Maryam, M. S., and Seyed, A. R. (2006). *Phase-based speech processing*. World Scientific.

Perrier, P., Ma, L., and Payan, Y. (2005). Modeling the production of vcv sequences via the inversion of a biomechanical model of the tongue. In *Proceedings of 9th European Conference on Speech Communication and Technology*, pages 1041–1044.

Perrier, P. and Ostry, D. J. (1996). The equilibrium point hypothesis and its application to speech motor control. *Journal of Speech and Hearing Research*, 39:365–378.

Perrier, P., Payan, Y., Zandipour, M., and Perkell, J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *Journal of the Acoustical Society of America*, 114(3):1582–1599.

Pisoni, D. B. and Remez, R. E. (2004). *The handbook of speech perception*. Oxford: Blackwell.

Pitton, J., Wang, K., and Juang, B.-H. (1996). Time-frequency analysis and auditory modeling for automatic recognition of speech. *Proceedings of the IEEE*, 84(9):1199 –1215.

Poo, G.-S. (1997). Large vocabulary mandarin final recognition based on two-level time-delay neural networks (tltdnn). *Speech Communication*, 22(1):17–24.

Rabaoui, A., Lachiri, Z., and Ellouze, N. (2004). Automatic environmental noise recognition. In *Industrial Technology, 2004. IEEE ICIT '04. 2004 IEEE International Conference on*, volume 3, pages 1670–1675 Vol. 3.

Rabiner, L. R. (1989). Tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital processing of speech signals*. Prentice-Hall.

Ragnier, M. and Allen, J. (2008). A method to identify noise-robust perceptual features: Application for consonant /t/. *Journal of the Acoustical Society of America*, 123(5):2801–2814.

Raj, B. and Stern, R. (2005). Missing-feature approaches in speech recognition. *Signal Processing Magazine, IEEE*, 22(5):101–116.

Ramirez, R. (2004). *Fast Fourier Transforms: Fundamentals and Concepts*. Prentice Hall.

Recasens, D. (1983). Place cues for nasal consonants with special reference to catalan. *Journal of the Acoustical Society of America*, 73(4):1346–1353.

Reddy, D. (1976). Speech recognition by machine: A review. *Proceedings of the IEEE*, 64(4):501–531.

Renals, S., Morgan, N., Bourlard, H., Cohen, M., and Franco, H. (1994). Connectionist probability estimators in hmm speech recognition. *IEEE Transactions on Speech and Audio Processing*, 12(1):161–174.

Richmond, K. (2009). Preliminary inversion mapping results with a new EMA corpus. In *Interspeech'09*, pages 2835–2838.

Robinson, T. and Fallside, F. (1991). A recurrent error propagation network speech recognition system. *Computer Speech and Language*, 5:259–274.

Rouat, J. (2008). Computational auditory scene analysis: Principles, algorithms, and applications (wang, d. and brown, g.j., eds.; 2006) [book review]. *IEEE Transactions on Neural Networks*, 19(1):199 –199.

Saenko, K., Livescu, K., Glass, J., and Darrell, T. (2005). Production domain modeling of pronunciation for visual speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal (ICASSP)*, volume V, pages 473– 476.

Saltzman, E. L. and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4):333–382.

Sankar, A., Kannan, A., Shahshahani, B., and Jackson, E. (2001). Task-specific adaptation of speech recognition models. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01*, pages 433–436.

Schafer, R. and Rabiner, L. (1975). Digital representations of speech signals. *Proceedings of the IEEE*, 63(4):662–677.

Scharenborg, O., Wan, V., and Moore, R. (2007). Towards capturing fine phonetic variation in speech using articulatory features. *Speech Communication*, 49(10-11):811–826.

Schroeder, M. R. (2004). *Computer Speech: Recognition, Compression, Synthesis*. Springer, 2nd edition.

Schroeter, J. and Sondhi, M. M. (1994). Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(1):133 –150.

Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673 –2681.

Schwarz, P., Matejaka, P., and Cernocky, J. (2006). Hierarchical structures of neural networks for phoneme recognition. In *Proceedings of ICASSP06, Toulouse, France*, pages 325–328.

Scott, S. and Johnsrude, I. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, 26(2):100–107.

Shimizu, Y., Kajita, S., Takeda, K., and Itakura, F. (2000). Speech recognition based on space diversity using distributed multi-microphone. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1747–1750 vol.3.

Sim, A. and Bao, P. (1998). *A Hybrid Speech Recognition Model based on HMM and Fuzzy PPM*. PhD thesis, The Hong Kong Polytechnic University.

Siniscalchi, S. and Lee, C. H. (2009). A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. *Speech Communication*, 51(11):1139–1153.

Sirigos, J., Fakotakis, N., and Kokkinakis, G. (2002). A hybrid syllable recognition system based on vowel spotting. *Speech Communication*, 38(3-4):427–440.

Siroux, J. and Gillet, D. (1985). A system for man-machine communication using speech. *Speech Communication*, 4(4):289–315.

Slaney, M. (1998). *Auditory toolbox*.

Slotine, J. J. E. and Li, W. (1991). *Applied Nonlinear Control*. Upper Saddle River, NJ: Prentice-Hall.

Sondhi, M. and Schroeter, J. (1987). A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(7):955–967.

Song, Q., Wu, Y., and Soh, Y. C. (2008). Robust adaptive gradient-descent training algorithm for recurrent neural networks in discrete time domain. *IEEE Transaction on Neural Networks*, 19(11):1841–1853.

Sroka, J. and Braida, L. (2005). Human and machine consonant recognition. *Speech Communication*, 45 (4):401423.

Steeneken, J. and Van Leeuwen, D. (1995). Multi-lingual assessment of speaker independent large vocabulary speech-recognition systems: the sqale project (speech recognition quality assessment for language engineering). In *Eurospeech '95, Proceedings of the Fourth European Conference on Speech Communication and Technology, Madrid, Spain.*

Stevens, K. (2002). Toward a model of lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, 111(4):1872–1891.

Stouten, V. and Hugo, V. h. (2009). Automatic voice onset time estimation from reassignment spectra. *Speech Communication*, 51:1194–1205.

Strom, N. (1997). Phoneme probability estimation with dynamic sparsely connected artificial neural networks. *The Free Speech Journal*, 5.

Suh, Y., Ji, M., and Kim, H. (2007). Probabilistic class histogram equalization for robust speech recognition. *Signal Processing Letters, IEEE*, 14(4):287 –290.

Titze, I. (1980). Comments on the myoelastic-aerodynamic theory of phonation. *Journal of Speech and Hearing Research*, 23(3):495–510.

Toda, T., Black, A., and Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model. *Speech Communication*, 50(3):215–227.

Trentin, E. and Gori, M. (2001). A survey of hybrid ann/hmm models for automatic speech recognition. *Neurocomputing*, 37(1-4):91–126.

van den Doel, K. and Ascher, U. (2008). Real-time numerical solution of webster's equation on a nonuniform grid. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1163 –1172.

Waibel, A. and Lee, K.-F. (1992). *Readings in Speech Recognition.* Morgan Kaufmann Publishers, Inc.

Wang, L. X. (1997). *A Course in Fuzzy Systems and Control*. NJ: Prentice-Hall.

Weenink, D. (2006). *Speaker adaptive vowel identification*. PhD thesis, University of Amsterdam.

Wells, J. C. (1997). SAMPA computer readable phonetic alphabet. In Gibbon, D., Moore, R., and Winski, R., editors, *Handbook of Standards and Resources for Spoken Language Systems*, chapter Part IV, section B. Berlin and New York: Mouton de Gruyter.

Windmann, S. and Haeb-Umbach, R. (2009). Parameter estimation of a state-space model of noise for robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1577 –1590.

Woodland, P. (1998). Speech recognition. In *Speech and Language Engineering - State of the Art (Ref. No. 1998/499), IEE Colloquium on*, pages 2/1–2/5.

Wrench, A. (1999). *The MOCHA-TIMIT articulatory database*.

Wrench, A. and Richmond, K. (2000). Continuous speech recognition using articulatory data. In *Proc. ICSLP 2000*, Beijing, China.

Wu, S., Er, M. J., and Gao, Y. (2001). A fast approach for automatic generation of fuzzy rules by generalized dynamic fuzzy neural networks. *IEEE Transactions on Fuzzy Systems*, 9(4):578 –594.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2006). *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.

Yousefian, N., Jalalvand, A., Ahmadi, P., and Analoui, M. (2008). Speech recognition with a competitive probabilistic radial basis neural network. In *Intelligent Systems, 2008. IS '08. 4th International IEEE Conference*, volume 1, pages 719–723.

Youssef, A., Badin, P., Bailly, G., and Heracleous, P. (2009). Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden markov models. pages 2255–2258.

Yu, H. J. and Oh, Y. H. (2000). A neural network for 500 word vocabulary word spotting using non-uniform units. *Neural Networks*, 13(6):681–688.

Zue, V. (2004). Fifty years of progress in speech understanding systems. *Journal of the Acoustical Society of America*, 116(4):2498–2498.

Zue, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., and Seneff, S. (1990). The voyager speech understanding system: preliminary development and evaluation. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 73–76 vol.1.