Multi-View Features in a DNN-CRF Model for Improved Sentence Unit Detection on English Broadcast News

Guangpu Huang*, Chenglin Xu†, Xiong Xiao*, Lei Xie†, Eng Siong Chng*‡, Haizhou Li*‡§
* Temasek Laboratories@NTU, Singapore
† Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, China
‡ School of Computer Engineering, Nanyang Technological University, Singapore
§ Institute for Infocomm Research, $A^{\star}$STAR, Singapore
E-mail: gphuang@ntu.edu.sg clxu@mail.nwpu.edu.cn xiaoxiong@ntu.edu.sg
lxie@nwpu.edu.cn aseschng@ntu.edu.sg hli@i2r.a-star.edu.sg

Abstract—**This paper presents a deep neural network-conditional random field (DNN-CRF) system with multi-view features for sentence unit detection on English broadcast news. We proposed a set of multi-view features extracted from the acoustic, articulatory, and linguistic domains, and used them together in the DNN-CRF model to predict the sentence boundaries. We tested the accuracy of the multi-view features on the standard NIST RT-04 English broadcast news speech data. Experiments show that the best system outperforms the state-of-the-art sentence unit detection system significantly by 13.2% absolute NIST sentence error rate reduction using the reference transcription. However, the performance gain is limited on the recognized transcription partly due to the high word error rate.**

I. Introduction

Automatic speech recognition (ASR) usually produces no punctuation in the output transcriptions, which makes it difficult for human to read and for text processing in downstream natural language processing (NLP) modules. So automatic sentence unit detection (SUD) serves as an important connection between ASR and NLP [1]. The goal of SUD is to detect the location of the sentence boundaries, or tag the word at the sentence end, in the raw text. There are mainly three types of full sentences, i.e., statement ('.'), question ('?'), and exclamation ('!') in English. At a lower level, speech boundary types such as comma (',') and incomplete ('...') marks are predicted [2]. They only account for the substitution error during the standard NIST evaluation. At a higher level, speech boundary types such as dialog and story boundaries are predicted [3], [4]. In this study, we focus on the SUD task of the three full sentences on English broadcast news (BN) data. BN is more challenging and interesting to segment than conversational speech, as the latter is usually shorter and has separate channels [5]. Though SUD is a tagging task, the existing NLP tagging tools are usually not directly applied to the speech transcriptions to detect the sentence boundaries. There are two main reasons [2], [6]. Firstly the spoken language uses a speech syntax that is different from the written text. For example, human speech often contains disfluencies such as repairs, false starts, and/or repetitions [2], [7]. Secondly the ASR output contains errors, especially for spontaneous speech. For example, BN data often contains both anchor speech and noisy speech between reporters and interviewees on the street. The errors in the ASR transcriptions will degrade the performance of the standard NLP taggers that are trained on written text. Moreover, spoken language contains knowledge sources that are not available in written text. These knowledge sources can help tackle the above difficulties in SUD. In the literature, conventional SUD systems use word-level and/or phone-level prosody features [8] together with the text in a statistical classifier such as, maximum entropy models[9], confusion networks [10], hidden Markov model (HMM), and conditional random field (CRF) [11] . The prosody features are usually heuristically defined. For example, Shriberg et al. extracted over 200 prosody features for sentence and topic segmentation in an HMM framework, and they obtained good results [12]. Ref. [2] combined the above prosody features with part-of-speech tags, and achieved state-of-the-art performance on English RT-04 data.

In recent years, there is trend to study human speech in both the production and the perception domains [13], [14]. Arora and Livescu described the concept of **multi-view features** for phonetic recognition in [13] . They described the two views with both the acoustic and the articulatory data, where the latter is available for a limited amount of training data, but not at test time. Their intuition is that the articulatory data provides additional information about the linguistic content in the two views. The articulatory dynamics describe the smooth and continuous movements in the vocal tract, which induce the acoustic variability of human speech. Compared to the acoustic features, the articulatory features are slow varying, and they are constrained by the physiological capacity of the human speech apparatus. This is especially useful when dealing with sentence boundary events [15]. In the absence of actual articulatory data, various methods have been proposed to predict the articulatory trajectories from acoustics. Recent success of DNN based acoustic-to-articulatory inversion, or speech inversion has also motivated us to exploit the potential of adding production knowledge in the SUD task. Currently the DNN based speech inversion system in [16] obtains mean square error rate as low as 0.83 mm on the x-y coordinates of tongue and lips on the MNGU0 corpus. Other techniques have also obtained around 1 mm error using HMM [14] and trajectory mixture density network (TMDN) [17], etc.

In this study, we apply the multi-view learning [13], where multiple views of data are available for training but possibly not for testing to the SUD task. We propose a set of multi-view features extracted from the acoustic, articulatory, and linguistic domains, and use them together in the DNN-CRF model to predict the sentence boundaries. The hypothesis is that the articulatory data may provide additional or complementary information about speech dynamics that are not available in other knowledge sources. We also implement a joint DNN-CRF model to segment English broadcast news into structured sentences [1]. This paper uses the multi-view features to the DNN-CRF, and compares the contributions of different features on speech inversion and sentence segmentation task. The proposed system outperformed the state-of-the-art system in [2] on English RT-04 speech corpus by 13.2% absolute NIST error rate on the reference transcription. The motivation is that the combined use of multiple production/perception knowledge sources on the linguistic content will improve the representation of sentence structures in spoken language. To our knowledge, this is the first time articulatory features have been used in SUD.

The rest of the paper is organized as follows. Section II describes the multiple knowledge sources and the DNN-CRF model for SUD. Section III describes the experimental setup and evaluation metrics. Section IV analyzes the results in comparison with the state-of-the-art SUD system. Section V concludes the paper.

II. Method

A. Multi-view Features

We extract the proposed multi-view features from the production, the perception, and the linguistic domains, which are closely related to the sentence structure in spoken speech. The basic multi-view features are consisted of 54 acoustic features (AcFs), 108 articulatory feature (ArFs), 162 prosody features (PFs), and 8 linguistic feature (LF), including 5-gram, part-of-speech (pos) tag, chunk (chk), and named entity (ne) tag, The four types of features and their relations to the spoken speech and the written text are shown in Fig. 1.

Prosodic View: The PFs used in [2], [12], and [18] consists of four groups of cues: pitch, energy, duration, and binary features such as speaker turn change or channel change. These studies have shown that the PFs are very effective in SUD on English BN data. We extracted the same set of 162 PFs as the baseline features. The PFs are derived with heuristically defined rules that are closely related to sentence boundaries. For example, pause duration before and after a word, stylized pitch/energy track, and speaker turn changes are primary cues for SUD [2], [12]. In the SUD system, the PFs of the word immediately preceding and following the boundary, or within a window of 20 frames (200 ms, empirically determined) before and after the boundary are extracted at each inter-word boundary. We used an open source automatic tool based on Praat to extract the prosodic features, as described in [12].

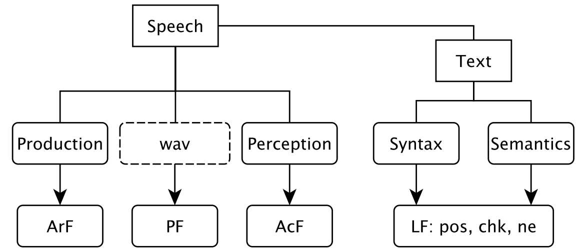

Fig. 1: Illustration of the multi-view speech features: acoustic feature (AcF), articulatory feature (ArF), prosody feature (PF), and linguistic feature (LF), including part-of-speech (pos) tag, chunk (chk) tag, and named entity (ne) tag.

Acoustic View: However, PFs are mostly static descriptions of the pitch/energy/duration features in the word intervals. They may under-represent the dynamics in spontaneous speech, e.g., BN data [19], [20]. To preserve such dynamics, we used auto-correlation method to track 6 basic AcFs: the energy, perceived pitch, and formants (F1 to F4) of the speech signal. The AcFs are estimated over 15 msec windows with a frame rate of 200 Hz, which also matches the electromagnetic midsagittal articulography (EMA) rate for speech inversion to extract the ArFs.

Articulatory View: Moreover, the dynamics in the AcFs are rooted in the dynamics in the production domain, i.e., the movements of the speech articulators. There is a close link between articulatory and acoustic representations of speech, where the syntactic structure of spoken sentences are directly constrained by the articulatory configurations[21], [22]. For example, a normal spoken sentence is usually less than 20 words long, as the lungs have limited capacity of air. In the proposed SUD system, we used a DNN inversion module to extract 12 basic ArFs from the acoustic speech signal and used them in the SUD task. The basic ArFs include the x-y coordinates of the 6 articulators, the upper lip, lower lip, lower incisor, tongue tip, tongue blade, and tongue dorsum. The hypothesis is that the articulatory dynamics are slow varying and provide additional production knowledge. The DNN was trained to estimate articulatory trajectories from input speech. The training data is a corpus of 460 English TIMIT sentences with EMA and microphone recordings. Previous studies have found that deep architectures give more accurate predictions than shallow ones, as the former has a higher capability to represent the mapping relationships between the input and the output. The best model obtained an average RMSE of 0.94

[1] Some of our results on the DNN-CRF model were reported in [18]

mm (with random initialization) on the MNGU0 test data [17], [23]. We used a similar DNN module to invert the ArFs, albeit much simpler features and structure, and obtained 0.99 mm accuracy on the same MNGU0 test data. We will elaborate more on the experimental settings and results in Section IV.

The sentence structure in spontaneous speech directly influences the continuity and dynamics of these features across word boundaries. Therefore, the derivatives of the above AcFs and ArFs are calculated to prepare a set of fixed-length feature vectors for DNN input in the proposed SUD system. We generate the word-level AcFs and ArFs from their previous frame-level measurements using two tiers of measurements. Tier one derivatives are the [*mean, std, max, min, median*] of the original features. Tier two derivatives are the [*mean, std, max, min*] of the *moving slope* and the 9^{th} order *polynomial fit* of the original features. The resulting AcF is $6 \times 18 = 108$ dimensional, while the ArF is $12 \times 18 = 216$ dimensional for each word interval.

Linguistic View: We used the textual, syntactic, and semantic features, namely the linguistic features (LFs) for SUD. The textual features, or the word sequences are the basic linguistic cues available either from the ground-truth or reference transcriptions (denoted as REF) or the ASR transcriptions (denoted as ASR). In fact, REF can be viewed as the ASR output with 100% word recognition rate. We used 5-grams as the textual features, i.e., $< w_{i-2}, wi-1, w_i, w_{i+1}, w_{i+2} >$, where w_i is the i^{th} word in the word sequence that will be tagged. Parsing uncovers the word relationships in the sentence, where the syntactic and semantic features are tightly connected to the sentence structure, i.e., the grammar in spoken language. [1]. Therefore, besides the textual transcriptions, we used the following three types of syntactic and semantic features: part-of-speech (POS) tags, chunks, and named entity tags. The POS tags and chunks have been shown to improve SUD performance on English data in the stat-of-the-art SUD system [2]. In addition, named entity tags are important cues for sentence boundary in BN data, where the reporters usually finishes their report with their names at the end of the news segment. In Section IV, we will show that the LFs can improve the SUD consistently. We used the existing SENNA tagger [24] [2] in the proposed SUD system.

B. Sentence Unit Detection Method

SUD is a classification task, i.e., the system needs to determine whether there is a sentence boundary label at each inter-word boundary. So the inter-word positions are labeled as either sentence unit (SU) or non-sentence unit (NS). Previously Liu et al. implemented the HMM system in [2], which combines the textural features in the n-gram model and the prosody features in the decision tree (DT). In particular, the observations consist of words and prosodic features. The prosody features are modeled as observation likelihoods attached to the n-gram states of the HMM [3], [5]. However, HMM has two main drawbacks [2]. First, the standard training methods maximize the joint probability of observed and hidden events, as opposed to the posterior probability of the correct hidden variable assignment given the observations, which is a criterion more closely related to classification performance. Second, the n-gram model makes it difficult to use features that are highly correlated, which would make robust estimation difficult. e.g., for ASR transcriptions. Liu et al. further implemented a CRF based SUD system to compete with the HMM system. CRF differs from HMM in that its training objective function is the conditional likelihood rather than the joint likelihood [25]. HMM does not maximize the posterior of the correct tags, while the CRF directly estimates the posterior boundary label probabilities. The conditional likelihood is closely related to the individual event posteriors for classification, which enables the CRF model to explicitly optimize discrimination of correct from incorrect event tags. Previous studies have shown that CRF can outperform the hidden-event LM baseline by 10% relative on the REF transcriptions and by up to 25% relative on the ASR transcriptions [26]. However, one drawback is that the CRF takes longer to train than the HMM, especially when the number of features becomes large.

In this study, we replaced DT with DNN to model the prosody features. The proposed SUD system is shown in Fig. 2. The input of the DNN is the prosodic/acoustic/articulatory features: PFs, AcFs, ArFs and the output is the posterior of the features. The LFs are used together with the DNN posterior as input to the linear-chain CRF for boundary detection. In other words, the boundary/non-boundary labels are decoded as the hidden events in the CRF model at each inter-word region across the speech transcripts. The DNN is trained to correctly classify the word boundaries into two classes, i.e. sentence unit (SU) or non-sentence unit (NS). In the input layer, the various features have the same length for each word, and they are fed to DNN. In the hidden layers, DNN learns the non-linear transformation in each hidden layer. The activation at layer l is defined as,

$$h_l = f_l(W_l h_{l-1} + b_l), \text{ for } 1 \leq l \leq L, \qquad (1)$$

where f_l is a sigmoid function, L is the total number of hidden layers, W_l is a weight matrix, and b_l is a bias vector. In the output layer, DNN adopts a softmax function to classify the boundary/non-boundary events given the input observations.

The DNN module calculates the posterior probabilities $P(E|F_k)$ of the SU event conditioned on the input features. They are combined with the linguistic features F_L to serve as observations in a linear-chain CRF model. The CRF model assigns a conditional probability distribution over the possible label sequences on a given training set, using the maximum likelihood criterion. In other words, the CRF integrates the multiple knowledge sources, and calculates the overall conditional probability distribution $P(E|O)$, where E represents the SU/NS labels sequences, and O represents the input observation sequences, F_L and $P(E|F_k)$ [2]. The most

[2] Online at http://ml.nec-labs.com/senna/

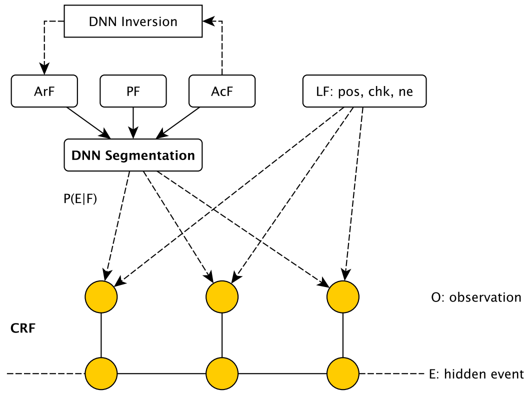

Fig. 2: Structure of the proposed DNN-CRF sentence unit detection system using the multi-view features as input.

like SU/NS sequence $\hat{E}$ is defined as,

$$\hat{E} = \arg\max_{E} P(E|O) \tag{2}$$

$$= \arg\max_{E} \frac{exp(\sum_k^K \lambda_k * F_k(E,O))}{\sum_E exp(\sum_k^K \lambda_k * F_k(E,O))}, \tag{3}$$

where k indicates the different features, each of which has an associated weight λ_k. For an input sequence O and an label sequence E, the posterior $F_k(E,O)$ is defined as,

$$F_k(E,O) = \sum_i f_k(E,O,i), \tag{4}$$

where i is an index over all the input positions. $f_k(E,O,i)$ is the feature function at position i over the label sequence and observation sequence. CRF uses a convex loss function which guarantees the convergence to a global optimum. The Viterbi algorithm is used to find the most likely label sequence. When $f_k(E,O,i) = f_k(E_{i-N},...,E_i,O_{i-M},...,O_i,i)$, an N-order linear-chain CRF, which models N $(E = E_{i-N},...,E_i)$ sequence labels and M $(O = O_{i-M},...,O_i)$ context features in the feature set, is formed. In practice, $N = 1$ and $M = 1$ are usually used because of the exponential increase of computational cost for higher N and M [2]. Details of the system training/test procedures are given in Section III-C.

III. Experiments

A. Data and Evaluation Metric for Speech Inversion

To extract the ArFs from the acoustic speech, we need a corpus with parallel articulatory and acoustic recording, like the MNGU0 corpus. The MNGU0 corpus consists of 1263 TIMIT sentences uttered by a single speaker. It contains ground-truth EMA recordings, microphone recordings, and phone-level transcriptions [17]. EMA is the most widely used articulography technique for creating parallel acoustic and articulator-position recordings. The electromagnetic transducer coils are glued to the vocal-tract articulators to record precise measurements of their positions. The 6 transducer coils are positioned at the upper lip (UL), lower lip (LL), lower incisor (LI), tongue tip (TT), tongue blade (TB), and tongue dorsum (TD). We used the 12-dimensional EMA measure as the raw ArFs, i.e., the x and y position of the 6 articulatory positions. The setting was also used in [16], [17]. The sampling frequency is 200 Hz. The dataset is partitioned into three sets: validation and testing sets comprising 63 utterances each, the training set consisting of the other 1137 utterances. We used the root mean square error (RMSE) to measure the accuracy of the predicted trajectories of every articulator i, where its RMSE is defined as,

$$RMSE_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(A_i(t) - \hat{A}_i(t))}, \tag{5}$$

where A_i is the ground-truth trajectory and $\hat{A}_i$ is the estimated trajectory of length T.

B. Data and Evaluation Metric for SUD

We evaluate the SUD system on the English corpus (LDC2004T12) from the standard NIST rich transcription (RT)-04 Fall SUD task in the DARPA EARS program. The released portion only contains the training set used in the NIST task. To reproduce the results in [2], we extract 2 hours BN data from the RT-04 set to use as the test set and the rest 18 hours as the training set. Table I shows the data structure of the English BN corpora in the training/testing set. About 8% of the inter-word positions are sentence boundaries, i.e., the sentence is on average 12 to 13 words long. The reference transcription (REF) is annotated according to the annotation guideline in [27]. The recognized transcription (ASR) is generated from our speech recognizer with a word error rate of 29.5%.

The DNN-CRF SUD model was trained with the ground-truth REF transcriptions. They were tested on both REF and ASR transcriptions to study the influence of ASR errors on the segmentation system. Several evaluation metrics have been used for SUD. In this study, we reported the precision, recall, F1-score, and the NIST SU error rate. The F1 score (or F-measure) has previously been used by Shriberg et al. [12] and Liu et al. [7] to compare SUD performance. F1 is defined as,

$$F1 = \frac{(1+\beta^2) * Recall * Precision}{\beta^2 * Recall + Precision}, \tag{6}$$

where Precision = TP/(TP+FP) and Recall = TP/(TP+FN). TP denotes the number of true positives, FP denotes false positives, FN denotes false negatives, and β corresponds to the relative weight/ratio of precision versus recall, β = Precision/Recall. The NIST SU error rated is calculated using

TABLE I: Data structure of the English broadcast news corpora in the training/testing set.

English RT-04		
Training	Number of words	169,842
	Number of sentence units (SUs)	13,182
Test	Number of words	13,993
	Number of sentence units (SUs)	1,197

the standard scoring tool *md-eval-v17.pl*, which calculates the average number of misclassified boundaries per reference boundary. For the ASR transcriptions, the scoring tool aligns the reference word string (REF) and the hypothesized word string (ASR) to minimize the word error rate [3]. Then the hypothesized SU events are mapped to the reference events using the word alignment information, and the unmatched events, i.e., insertions and deletions are counted. The NIST SU error rate is defined as,

$$NIST\ SU\ error\ rate = \frac{insertion + deletion}{total\ number\ of\ SUs}. \qquad (7)$$

C. System Implementations

The multi-view features, or more precisely their DNN posteriors, are combined with the linguistic features in the DNN-CRF model. We trained the DNN, both the inversion and segmentation modules, in a supervised greedy layer-wise manner. We started with 1-hidden layer NN that mapped the PFs/AcFs/ArFs or their combinations to the SU/NS posterior probabilities, SU: sentence unit, NS: non-sentence unit. Then the output of the single layer DNN is used as the input to a second 1-hidden layer NN to map to the SU/NS posteriors. The procedure can be repeated with unlimited number of hidden layers. For the sentence boundary classification/tagging problem, we show that a 3-hidden layer DNN reaches the best performance, and adding more hidden layers does not further improve the performance. DNN training is implemented by stochastic gradient descent algorithm. Since there are limited date in the SUD task, we empirically set the L2 weight decay to a small value, 0.00001, to prevent over-fitting. The same setting was used in [18]. DNN training stops when the performance improvement is less than 0.002% on the validation data (part of the training set). The CRF++ toolkit [4] is used to implement the CRF model in this paper [28]. The DNN posteriors are quantized to 6 bins: $[0, 0.1], (0.1, 0.3], (0.3, 0.5], (0.5, 0.7], (0.7, 0.9], (0.9, 1]$, as the toolkit only handles discrete features.

IV. Results

A. Results on Speech Inversion

Table. II shows the detailed RMSE of the 12 dimension EMA data, or the raw ArFs, using the DNN inversion module and the MNGU0 corpus. The input to the inversion module consists of 5 selected AcFs, i.e., the energy, pitch, and formant tracks (F1 to F3), where [16] uses 40 frequency warped line spectral frequencies (LSFs) and an energy gain. Formant 4 (F4) is not used as we found that it degrades the RMSE by about 0.05 mm. The context window length is 5. The DNN inversion module has 3 hidden layers, each with 100 hidden nodes. The output has 12 nodes, which are the x-y coordinates of the 6 articulators, the upper lip (UL), lower lip (LL), lower incisor (LI), tongue tip (TT), tongue blade (TB), and tongue dorsum (TD). The same outputs are used [16]. We also used a low-pass filter (LPF) with a cut-off frequency at 15 Hz to smooth the inverted ArFs by eliminating unrealistic or abrupt articulatory movements. The reason of using a low frequency is that the articulatory trajectories from the acoustic-to-articulatory mapping are already smooth. It is not necessary to use a high cut-off frequency in the low-pass filter to smooth out abrupt trajectories The smoothed ArFs are used in the same way as the AcFs to produce word-level feature vectors for sentence segmentation.

We obtained RMSE of 0.99 mm on the same MNGU0 test data, where [16] obtained 0.93 mm on DNN (and reduced to 0.88 mm with pre-training), and [17] obtained 0.99 mm on TMDN. However, the selected 5-dimensional AcFs are able to retrieve the ArFs with similar precision as the 40-dimensional LSFs, as used in [16], [17]. The AcFs are directly related to the articulatory trajectories. For example, one region in the articulatory space, 'fibers', could correspond to a single point in the acoustic parameter space [14], [17]. The AcFs also demonstrate different error patterns from the LSFs. They obtain higher error on the y-coordinates of the TT, TB, and TD than their x-coordinates. This is expected since these articulators have more abrupt movements on the y-axis, i.e., high-low tongue positions. For example, the tongue tip movement during the production of dental plosives (e.g., /t, d/) and the jaw movement during the production of the low vowels (e.g., /aa, ow/) are more abrupt than the other articulators.

TABLE II: Detailed RMSE (mm) of the DNN inversion module on the MNGU0 test set. The 12 articulatory dimensions include the x-y coordinates of the 6 articulators, the upper lip (UL), lower lip (LL), lower incisor (LI), tongue tip (TT), tongue blade (TB), and tongue dorsum (TD).

	TMDN [17]	LSF-DNN [16]	AcF-DNN	
			raw	LPF
UL_x	0.32	-	0.34	0.34
UL_y	0.49	-	0.47	0.44
LL_x	0.64	-	0.55	0.54
LL_y	1.18	-	1.04	0.99
LI_x	0.57	-	0.56	0.55
LI_y	0.75	-	0.82	0.79
TT_x	1.36	-	0.74	0.72
TT_y	1.28	-	2.49	2.39
TB_x	1.34	-	0.64	0.62
TB_y	1.24	-	2.37	2.29
TD_x	1.22	-	0.71	0.67
TD_y	1.57	-	1.93	1.86
Average (mm)	0.99	0.93	**1.05**	**0.99**

B. Results on SUD

Table III summarizes the sentence unit detection results (precision, recall, F1, and NIST error rate) on the broadcast news (BN) RT-04 corpus. Two types of transcriptions: human-generated reference transcriptions (REF) and automatic speech recognition output (ASR) are used in the different feature-system settings: The results of the HMM, CRF, and DT-CRF systems are extracted from [2], which used prosody features and linguistic features (without the named entity tags). The

[3] Online at http://www.itl.nist.gov/iad/894.01/tests/rt/2004-fall/

[4] Online at http://crfpp.googlecode.com/svn/trunk/doc/index.html

DT-CRF used a C4.5 tree structure, where the DT posteriors are combined with the lexical features in the CRF model [2]. Using the PFs and the LFs on the DT-CRF model, Liu et al. achieved 43.1% NIST SU error rate on the English RT-04 BN REF data, and 55.6% on the ASR data [2]. The difference between the test condition in [2] and ours is that we split the available RT-04 data for training and testing, which is part of the full RT-04 data as used by [2] in their evaluation. In addition, the word error rate in the ASR transcripts of [2] is 11.7%, much lower than 29.5% our ASR transcripts.

Using the same feature input, DNN-CRF outperforms the DT-CRF model of [2], with 7.2% and 2.3% absolute NIST SU error reduction on the REF and ASR transcriptions. With the multi-view features, DNN-CRF: multi-view features further reduce the NIST SU error by 6.0% absolute rate on the REF transcriptions. There is little performance gain from the multi-view features in the ASR condition. Yet the multi-view DNN-CRF system shows slightly higher precision but lower recall rate than the previous system. Both the true positive and the false alarm rates are much lower in the new system. One issue is the high word error rate in the recognition output, 29.5%, which affects the features extraction for each word interval. Word errors may propagate in the LFs through POS tagging, chunking, and named entity tagging, where the LFs may be extracted with imperfect word alignments. For the ASR transcripts, the precision is over 90%, which indicates that there are more missing SUs compared to the REF condition. The recall is low for all three systems, and the final F1 value is not much improved with new features.

Table IV summarizes the NIST SU error rate of different feature combinations in the DNN-CRF model. The multi-view features achieve the best NIST error rate, 29.9%, on the English RT-04 corpus. Results in Table IV also show that the PFs, AcFs, and ArFs are complementary to each other in the SUD task. One drawback is that the data size becomes large for the multi-view features, resulting in long DNN training time. In addition, the best performance by the multi-view feature, 29.9%, is similar to that of $PF + AcF$, 30.4%, which may be due to the fact that the ArFs are projected/inverted from the AcFs. We may need to find a more efficient way to exploit the advantage of the articulatory dynamics. Furthermore, we also need to tune the DNN posterior module on the limited data to optimize the performance, e.g., through pre-training, as the results tend to vary with random initialization.

TABLE IV: NIST SU error rate (%) of the DNN-CRF based sentence segmentation using different feature combinations on the English REF transcriptions.

	Feature (dimension)	DNN-CRF
Prosodic view	PF (162)	57.1
Acoustic view	AcF (108)	39.6
	PF + AcF	**30.4**
Articulatory view	ArF (216)	39.6
	PF + ArF	31.2
Multi-view	PF + AcF + ArF	**29.9**

C. Demo of Multi-View Features on English Sentence

Compared with the state-of-the-art SUD system in [2], the combined use of the prosody, the articulatory, the acoustic, and the linguistic features in the proposed DNN-CRF model results in improved SUD performance. The performance gain comes from two aspects. First, the proposed multi-view features are from the production, the perception, and the linguistic domains. They are useful when representing the speech dynamics at the sentence level, and they are complementary to each other in SUD. Second, the DNN-CRF model can effectively leverage the sequential information in the SU tagging problem [18], [28].

Fig. 3 shows the LFs (REF text, pos[5], chunk[6], & named entity[7]) and AcFs (intensity, pitch, and the first two formants) in the word-intervals (boundaries marked by dashed lines). The SU/NS labels are shown at the bottom layer in panel (1) of Fig. 3. At the sentence boundary, there is usually silence, and discontinuity of pitch value, as shown at the last word interval in panel (3) of Fig. 3 . However, there maybe overlaps among the groups of features. For example, [29] has shown that the stress patterns are correlated with the adjective-noun compounds in English speech. We have tried to de-correlate the features using principle component analysis, and the SUD performance degrades slightly on the reduced feature sets. This issue would be addressed in our future studies.

Moreover, many of the word boundaries are clearly visible in the AcF contours. So the proposed multi-view features and DNN-CRF model can also be used to detect other speech segment boundaries, such as word, paragraph, and story boundaries.

V. Discussion and Conclusion

Sentence unit detection adds punctuation marks to the ASR output, making it easier for human to read and for downstream NLP modules. It serves as an important connection between the ASR and the NLP modules.

In fact, SUD has already been widely applied at the ASR output for many languages besides English [2], e.g., Arabic [30], Chinese [31], Portuguese [32], etc. Though there are many parsing and segmentation technique in NLP that deals with pure text, the ASR output text contains no capitalizations and limited paragraph structures. Yet spontaneous speech contains multiple information sources that are not available in pure NLP text. These information sources can be used to predict the sentence boundaries in the running speech and its ASR transcripts. In this paper, we describe a DNN-CRF approach with multi-view features for sentence unit detection in English broadcast news. There are two main contributions and observations in this study. First, we extracte

[5] pos: JJ - adjective, TO - to, VB - Verb/base form, PRP - personal pronoun, NN - noun, singular or mass, VBG - verb/gerund or present participle, DT - determiner, IN - preposition, VBD - verb, past tense.

[6] chunk: S - beginning, ADJP - adjective phrase, VB - verb chunk, NP - noun chunk.

[7] named entity: E - entity, O - other. The su tags: NS - non-sentence, SU - sentence unit.

TABLE III: Performance of different feature-model combinations for the sentence detection task (precision, recall, F1, and NIST SU error rate) on the broadcast news (BN) RT-04 corpus.

	REF				ASR			
	P	R	F1	NIST SU error	P	R	F1	NIST SU error
HMM [2]	-	-	-	52.23	-	-	-	60.64
CRF [2]	-	-	-	49.88	-	-	-	58.21
DT-CRF [2]: PF+LF	81.4	73.9	77.4	43.1	90.6	49.5	64.0	55.6
DNN-CRF [18]: PF+LF	85.9	76.7	81.0	35.9	95.0	49.3	64.9	53.3
DNN-CRF: multi-view features	88.41	80.1	84.1	**29.9**	96.9	48.9	65.0	**52.6**

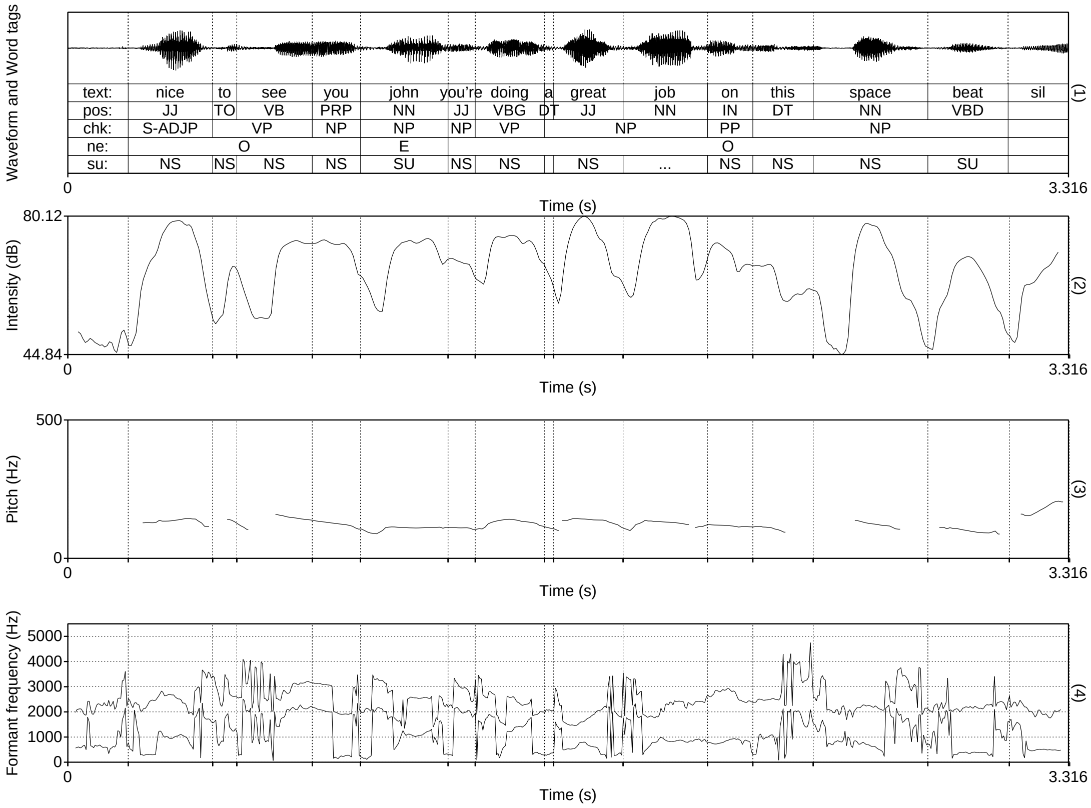

Fig. 3: The waveform with the selected acoustic and linguistic features in the RT-04 sentences $ee970703$: $nicetoseeyoujohnyou'redoingagreatjobonthisspacebeat$ panel (1) illustrates the linguistic features (REF text, pos, chunk (chk), named entity (ne) and the SU/NS tags. panel (2) illustrates the intensity contour; panel (3) illustrates the pitch contour; panel (4) illustrates the first two formants.

a set of multi-view features: prosody, acoustic, articulatory, and linguistic features, and use them together in the DNN-CRF model to predict the sentence boundaries. Experiments show that the proposed multi-view feature outperforms the best SUD system in the literature by 13.2% absolute NIST SU error rate. We also show that they played complementary roles during prediction, and achieved the best performance when used together. Second, we show that DNN is more accurate and more robust than DT model when modeling the speech features, both on the ground-truth REF transcripts and on the ASR transcripts. However, the accuracy of sentence detection is subject to the word-recognition accuracy in the ASR transcriptions. When applying NLP modules such as POS tagger and named entity recognizer, the word errors propagate to the modules and degrade the subsequent system performance. On the ASR transcripts, the DNN-CRF obtains less performance gain than on the REF transcripts. Our current work include testing the SUD module on different ASR error rates. Another issue in the SUD task is that there are more non-SUs than SUs in the training/test data, i.e., imbalanced data distribution [7]. So we are also interested in trying out different sampling/bagging methods to tune the DNN-CRF model.

References

[1] B. Favre, R. Grishman, D. Hillard, H. Ji, D. Hakkani-Tur, and M. Ostendorf, "Punctuating speech for information extraction," in *IEEE International Conference*

on Acoustics, Speech and Signal Processing, ICASSP 2008., 2008, pp. 5013–5016.

[2] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526 –1540, 2006.

[3] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," *Proc. ICSLP 2002*, pp. 2037–2040, 2002.

[4] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "Toward joint segmentation and classification of dialog acts in multiparty meetings," *Lecture comments in Computer Science (including subseries Lecture comments in Artificial Intelligence and Lecture comments in Bioinformatics)*, vol. 3869 LNCS, pp. 187–193, 2006.

[5] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," *Proc. of the ISCA Workshop: ASR-2000*, pp. 228–235, 2000.

[6] D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman, "Measuring the readability of automatic speech-to-text transcripts," *Proc. Eurospeech*, pp. 1585–1588, 2003.

[7] Y. Liu, N. Chawla, M. Harper, E. Shriberg, and A. Stolcke, "A study in machine learning from imbalanced data for sentence boundary detection in speech," *Computer Speech and Language*, vol. 20, no. 4, pp. 468–494, 2006.

[8] J. Gomez and M. Calvo, "Improvements on automatic speech segmentation at the phonetic level," *Lecture comments in Computer Science (including subseries Lecture comments in Artificial Intelligence and Lecture comments in Bioinformatics)*, vol. 7042 LNCS, pp. 557–564, 2011.

[9] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," *Proc. ICSLP*, pp. 917–920, 2002.

[10] D. Hillard, M. Ostendorf, A. Stolcke, Y. Liu, and E. Shriberg, "Improving automatic sentence boundary detection with confusion networks," *Proc. HLT-NAACL*, pp. 69–72, 2004.

[11] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," *Proc. 18th International Conf. on Machine Learning*, pp. 282–289, 2001.

[12] E. Shriberg, A. Stolcke, and D. Hakkani-tür Hakkani-tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, 2000.

[13] R. Arora and K. Livescu, "Multi-view cca-based acoustic features for phonetic recognition across speakers and domains," in *ICASSP*, 2013, pp. 7135–7139.

[14] T. Toda, A. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[15] G. Tur, A. Stolcke, L. Voss, S. Peters, D. Hakkani-Tur, J. Dowding, B. Favre, R. Fernandez, M. Frampton, M. Frandsen, C. Frederickson, M. Graciarena, D. Kintzing, K. Leveque, S. Mason, J. Niekrasz, M. Purver, K. Riedhammer, E. Shriberg, J. Tien, D. Vergyri, and F. Yang, "The calo meeting assistant system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1601–1611, 2010.

[16] B. Uria, I. Murray, S. Renals, and K. Richmond., "Deep architectures for articulatory inversion," in *Proc. Interspeech*, 2012.

[17] K. Richmond, "Preliminary inversion mapping results with a new ema corpus," in *Proc. Interspeech*, 2009, pp. 2835 –2838.

[18] C. Xu, L. Xie, G. Huang, X. Xiao, E. S. Chng, and H. Li, "A deep neural network approach for sentence boundary detection in broadcast news," in *Proc. Interspeech*, 2014.

[19] M. Gregory, M. Johnson, and E. Charniak, "Sentence-internal prosody does not help parsing the way punctuation does," *Proc. NAACL*, pp. 81–88, 2004.

[20] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 2956–2970, 1991.

[21] S. Millotte, A. Rene, R. Wales, and A. Christophe, "Phonological phrase boundaries constrain the online syntactic analysis of spoken sentences," *Journal of Experimental Psychology: Learning Memory and Cognition*, vol. 34, no. 4, pp. 874–885, 2008.

[22] D. Fogerty and D. Kewley-Port, "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility.," *The Journal of the Acoustical Society of America*, vol. 126, no. 2, pp. 847–857, 2009.

[23] B. Uria, S. Renals, and K. Richmond., "A deep neural network for acoustic-articulatory speech inversion," in *In NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[24] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa., "Natural language processing (almost) from scratch," *Journal of Machine Learning Research (JMLR)*, 2011.

[25] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," *Proceedings of CoNLL-2003*, pp. 188–191, 2003.

[26] A. Stolcke, "Srilm - an extensible language modeling toolkit," *Proc. ICSLP*, vol. 2, pp. 901–904, 2002.

[27] S. Strassel, "Simple metadata annotation specification," *Simple Metadata Annotation Specification V6.2*, 2004.

[28] T. Oba, T. Hori, and A. Nakamura, "Sentence boundary detection using sequential dependency analysis combined with crf-based chunking," vol. 3, 2006, pp. 1153–1156.

[29] T. Morrill, “Acoustic correlates of stress in english adjective-noun compounds,” *Language and Speech*, vol. 55, no. 2, pp. 167–201, 2012.

[30] A. Al-Subaihin, H. Al-Khalifa, and A. Al-Salman, “Sentence boundary detection in colloquial arabic text: a preliminary result,” 2011, pp. 30–32.

[31] C. Xu, L. Xie, and Z. Fu, “Sentence bounary detection in chinese broadcast news using conditional random fields and prosodic features,” in *The 2nd IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP2014)*, 2014.

[32] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, “Recovering capitalization and punctuation marks for automatic speech recognition: case study for portuguese broadcast news,” *Speech Communication*, vol. 50, no. 10, pp. 847–862, 2008.