Sentence Boundary Detection in Arabic Speech

Amit Srivastava, Francis Kubala

BBN Technologies
Cambridge, MA 02138
{asrivast, fkubala}@bbn.com

Abstract

This paper presents an automatic system to detect sentence boundaries in speech recognition transcripts. Two systems were developed that use independent sources of information. One is a linguistic system that uses linguistic features in a statistical language model while the other is an acoustic system that uses prosodic features in a feed-forward neural network model. A third system was developed that combines the scores from the acoustic and the linguistic systems in a Maximum-Likelihood framework. All systems outlined in this paper are essentially language-independent but all our experiments were conducted on the Arabic Broadcast News speech recognition transcripts. Our experiments show that while the acoustic system outperforms the linguistic system, the combined system achieves the best performance at detecting sentence boundaries.

1. Introduction

Speech-to-Text systems today produce transcriptions that differ in many ways from the normal written forms of a language. In almost every case, these differences result in a degraded transcript in terms of human readability. Most of today's systems do not hypothesize sentence boundaries, producing a monolithic transcript that is difficult to read. Absence of sentence boundaries also makes the speech recognition transcripts less useful to most Natural Language Processing systems. Sentences are the preferred basic units of most natural language understanding systems such as the Natural Language Parsers and the Information Retrieval systems.

Speech transcripts differ from text articles since they lack case, punctuation, and other structural cues. With the absence of such structure, it becomes very difficult to hypothesize sentences in speech transcripts. However, speech transcripts have one advantage in their favor. It has been shown that prosody has a strong correlation with discourse structure [4]. The availability of audio with the transcripts makes it possible to use prosody for detecting sentences. Prosody can make a big impact since pause and emphatic stress are part of the signature for many discourse structures like sentences.

Textual cues like word identity, word sequence, and part-of-speech, have been commonly used to build statistical language models on sentence beginnings and ends for sentence boundary detection. These language models rely on the fact that certain words are more likely to begin or end a sentence than others. Prosody has also been effectively used in detecting sentences in speech. Christensen, Gotoh and Renals [1], and Kim and Woodland [3], use pause duration, pitch, phoneme-duration, and energy to hypothesize sentences in English Broadcast News (BN) speech transcripts. Stolcke and Shriberg [5] have also used prosodic features, such as pitch, and pause in a decision-tree based system to detect sentences in English Conversational Telephone Speech (CTS) transcripts. Huang and Zweig [2] use lexical features and pause durations in a Maximum Entropy model for punctuation annotation in English CTS transcripts. However, most of the ongoing research has focused on the English language.

In this paper, we present our approach to sentence boundary detection in Section 2. In Section 3, we describe the experimental setup and the evaluation paradigm. In Section 4, experimental results on Arabic BN speech transcripts will be presented and discussed. Finally, we conclude the paper in Section 5.

2. Our Approach

Figure 1 shows the conceptual diagram of our approach. The output of our Arabic Audio Indexer system [7], with automatic speech recognition, speaker change detection, and named-entity extraction is funneled into the two subsystems, the *Acoustic* subsystem and the *Linguistic* subsystem. Each of these subsystems uses a different knowledge source and produces scores, which are estimates of the likelihood of each boundary class, *sentence-boundary (SB)*, or *no-sentence-boundary (NSB)*. The outputs from the two subsystems are used by the *Combined* system to produce the final hypothesis.

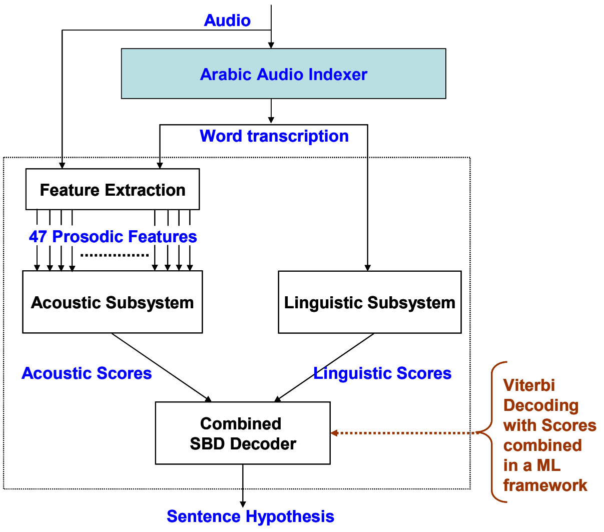


Figure 1: Conceptual diagram of our sentence boundary detection approach

Our sentence boundary detection system is trained on transcripts that have manually annotated words, speaker turns, and named-entities. During decoding, we use the output of our

Audio Indexer system with automatically hypothesized words, speaker turns, and named-entities. A number of prosodic features are extracted from the audio based on the word boundaries in the speech transcript. The *Acoustic* subsystem uses these prosodic features The *Linguistic* subsystem uses the hypothesized words and named-entities within each automatic speaker turn. The likelihood scores from each of the subsystems are then combined in a *Combined* SBD decoder to produce the final hypothesis of sentence boundaries.

We will describe the feature extraction component and the SBD subsystems in some detail and then describe the combination approach. Each SBD subsystem produces likelihood scores that can be used to derive a subsystem-level hypothesis. We analyzed the performance of each of these subsystems together with the performance of the combination stage.

2.1. Prosodic Feature Extraction

Figure 2 shows the conceptual diagram of our feature extraction method. We define putative boundaries, PB, as the intervals between words in the speech transcripts. The SBD system can hypothesize sentence boundaries in each of these PB. We extract a variety of prosodic features from the audio at each PB.

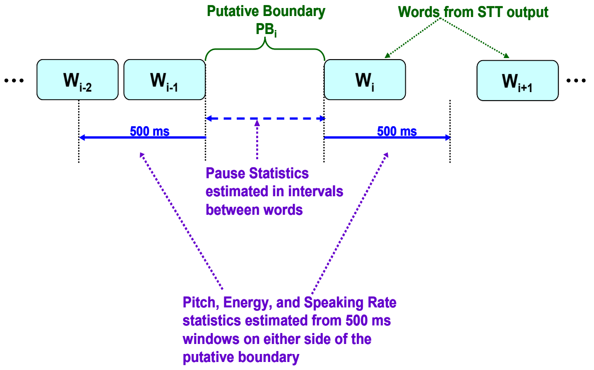


Figure 2: Prosodic feature extraction method

Pause statistics are estimated from the intervals between the words, which may be absent altogether when there is no pause between the hypothesized words. Pitch, Energy, and Speaking Rate statistics are estimated from 500 ms windows on either side of PB$_i$ extending from its edges. The estimated prosodic features are functions of these statistics.

We estimated four types of prosodic features in a total of 47 features. Table 1 shows the distribution of the prosodic features estimated as well as the types of statistics used to estimate these features. There were 9 *Pause* related features, 2 *Speaking Rate* features, 6 *Energy* based features, and 30 *Pitch* related features estimated at the putative boundaries. The prosodic features estimated during this effort were a diverse set of features with some being continuous, others being discrete, and yet others were Boolean features.

Feature Type	Feature Description	#Features
Pause	•Pause Duration •Pause Attribute (Filler, Breath, etc.) •Time since last Pause •Normalized Pause Duration	9
Speaking Rate	•Absolute Value •Difference across Putative Boundary	2
Energy	•Absolute Values •Difference across Putative Boundary •First Difference of Energy	6
Pitch	•Discontinuous Chains in Voiced Regions •Interpolated Continuous Pitch •First-Order Pitch Differences •Tone Estimates (Legendre Polynomials)	30

Table 1: Distribution of the prosodic feature types in our system

2.2. Acoustic Subsystem

The *Acoustic* subsystem consists of a 2-layer feed-forward Neural Network (NN) that was trained on the 47 prosodic features mentioned in Section 2.1. The NN system was chosen for its ability to determine the complex function that maps the diverse set of input prosodic features into the output boundary classes. The network configuration was 47 input nodes, 4000 hidden nodes, and 2 output nodes. Standard back-propagation training was applied with the Minimum Squared-Error (MSE) criterion.

Each putative boundary PB$_i$ (Figure 2) is associated with a boundary class c$_i$, which is one of *sentence-boundary (SB)*, or *no-sentence-boundary (NSB)*, and a 47-dimensional prosodic feature vector f$_i$. The *Acoustic* subsystem hypothesizes sentence boundaries by comparing the output of the NN against an empirical threshold as shown in the following equation:

$$c_i = \begin{cases} \mathit{sentence-boundary} & \text{if } \phi(f_i) \geq \tau \\ \mathit{no-sentence-boundary} & \text{if } \phi(f_i) < \tau \end{cases} \qquad (1)$$

where

- $\phi(f_i)$ is the NN output for the input prosodic feature vector f$_i$
- τ is an empirical threshold

2.3. Linguistic Subsystem

The *Linguistic* subsystem consists of a trigram language model that is very similar in spirit to the Finite State Model described in [1]. We used the BBN IdentiFinder system to convert names into name-class tokens, e.g., the words in the name “حسني مبارك” were converted into the tokens “*PERSON* *PERSON*”. The name-class mapping was applied to the training transcripts as well as to the speech recognition output. This mapping helps in reducing the effective lexicon in our language model. The need for a smaller lexicon was a constraint of our implementation during the course of this early work.

Following the mapping, each word w$_i$ has an associated boundary class c$_i$, which is one of *SB* or *NSB*. Word w$_i$ starts a sentence if the associated boundary class c$_i$ is *SB*. Sentence boundaries are hypothesized by the *Linguistic* subsystem using the Viterbi algorithm for finding the optimal sequence of boundary classes $\hat{c}_1^N$, given the sequence of words w_1^N in a speaker turn containing N words, as shown in the following equation:

$$\hat{c}_1^N = \underset{c_1^N}{\operatorname{argmax}}\, p^L\left(w_1^N, c_1^N\right). \qquad (2)$$

$$\approx \underset{c_1^N}{\operatorname{argmax}} \prod_{i=1}^{N} p^L\left(w_i / w_{i-1}, w_{i-2}, c_i\right)^{\lambda} * p^L\left(c_i / w_{i-1}, w_{i-2}, c_{i-1}\right) \qquad (3)$$

λ, here, is an exponential smoothing factor. The superscript p^L (.) denotes *Linguistic* likelihoods.

2.4. Combined System

In the *Acoustic* subsystem, the NN produces scores, $\phi(f_i)$ that are MSE estimates of the posterior probability of the boundary classes. These scores are transformed into likelihoods by scaling them with the prior probability, $P(c_i)$, of the boundary classes, estimated from the same training data. Thus, the *Acoustic* subsystem produces the acoustic likelihood, p^A, as follows:

$$p^A\left(f_i / c_i\right) = \frac{\phi(f_i)}{P(c_i)} \qquad (4)$$

The *Combined* system combines the scores from the two subsystems in a Maximum-Likelihood framework. The optimal sequence of boundary classes $\hat{c}_1^N$, given the sequence of words w_1^N, and the sequence of prosodic feature vectors f_1^N in a speaker turn with N words, is determined using the Viterbi algorithm as follows:

$$\hat{c}_1^N = \underset{c_1^N}{\operatorname{argmax}}\, p^A\left(f_1^N, c_1^N\right)^{\alpha} * p^L\left(w_1^N, c_1^N\right) \qquad (5)$$

$$\approx \underset{c_1^N}{\operatorname{argmax}} \prod_{i=1}^{N} p^A\left(f_i / c_i\right)^{\alpha} * p^L\left(w_i / w_{i-1}, w_{i-2}, c_i\right)^{\beta} * p^L\left(c_i / w_{i-1}, w_{i-2}, c_{i-1}\right) \qquad (6)$$

where α , β are exponential smoothing factors

3. Experimental Setup

We used Broadcast News acoustic data consisting of approximately 57 hours of spoken Arabic, transcribed without diacritic markings, from Egyptian and Syrian broadcast radio and TV, and the Al-Jazeera TV network for our experiments [7]. The 57 hours of acoustic data were split into a training set and a test set. The most recent 6.5 hours of data was chosen as the test set, while the rest was used for training. The corresponding transcriptions in this corpus contain sentence boundary annotation. The data used in our experiments is summarized in Table 2.

The neural network in the *Acoustic* subsystem was trained on 320K prosodic feature vectors estimated from the training data. Approximately 320K words, from the same acoustic training transcription, were used to train the language model for the *Linguistic* subsystem. The training data transcriptions contained manually annotated names of people, locations, and organizations. These were used for the name-class mapping prior to building the language model in the *Linguistic* subsystem, as well as to train the statistical models for the BBN IdentiFinder system [6].

Set	# hours	# words	# sentences
Training	50.5	320K	12K
Test	6.5	40K	1.6K
Total	57.0	360K	13.6K

Table 2: Experimental data summary

All systems were evaluated on the output of our Arabic Speech-to-Text (STT) system.wirh approximately 19% word error rate (WER). Speaker turns were defined using the automatic speaker change detection system and sentence boundaries were hypothesized within each speaker turn. The BBN IdentiFinder system, performing at 17% Slot Error Rate, was used to automatically identify the names in the STT output for the name-class mapping.

We generated the reference by mapping the sentence boundaries, by time, from the reference transcriptions to the nearest putative boundary in the STT output for each episode in the test set. The mapped STT output was now considered as the reference for the performance evaluation. During scoring, the system-output sentence boundaries were compared against the mapped-reference sentence boundaries in the STT output. We used the *Detection Error Rate* (DER) as our primary evaluation metric. DER is the sum of *False Acceptance Rate (FA)* and the *False Rejection Rate (FR)* and is similar to the Slot Error Rate metric proposed in [8]. A False Acceptance error is counted when the system hypothesizes a *sentence-boundary* when there is none. A False Rejection errors is counted when the system misses a true *sentence-boundary*. Systems were compared at the equal-error operating point to allow comparison of different systems under similar operating conditions.

4. Results

Table 3 shows the performance of the 3 systems, *Acoustic*, *Linguistic*, and *Combined*.

System	%FA	%FR	%DER
Acoustic	27.88	27.82	**55.70**
Linguistic	36.90	39.60	**76.50**
Combined	25.13	25.25	**50.38**

Table 3: Performance of the SBD systems on STT output

The results show that the *Acoustic* subsystem outperforms the *Linguistic* subsystem by a wide margin. It is not surprising that the *Linguistic* subsystem is weak at detecting sentence boundaries given that the only training data available for building the language model consists of the acoustic transcriptions, which is quite miniscule. The *Combined* system produces a substantial gain even when combining two very different systems, indicative of its ability to effectively combine different knowledge sources like the prosodic information and the linguistic information to detect sentence boundaries.

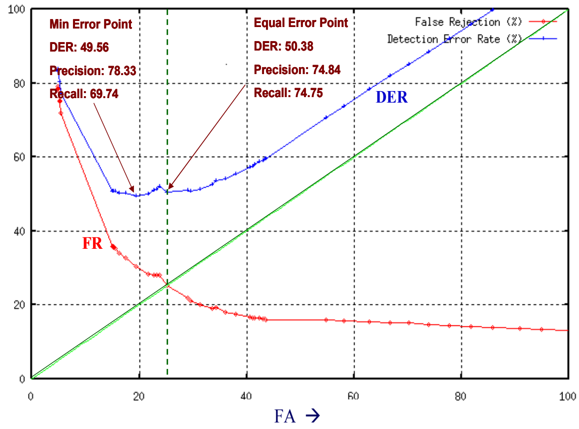
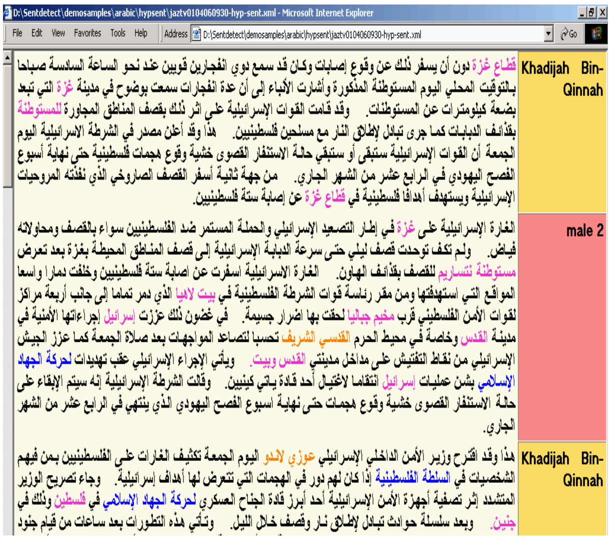


Figure 3: DET plot for the *Combined* system on Arabic STT output

We tuned the parameters, α , β , in Equation (6) to analyze the performance of the *Combined* system in a DET plot, which is a plot of FR versus FA, as well as a plot of the DER versus FA, as shown in Figure 3. The diagonal line across the diagram can be used to identify the equal-error points on the two plots. We can see from the DET plot that the equal-error operating point, 50.38% DER, is very close to the minimum-error operating point, 49.56% DER. This indicates that our system is not biased towards one kind of error even though the *no-sentence-boundary* class dominates the boundary classes in the training and test transcriptions.


D:\Sentdetect\demosamples\arabic\hypsent\jaztv0104060930-hyp-sent.xml - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Address
D:\Sentdetect\demosamples\arabic\hypsent\jaztv0104060930-hyp-sent.xml
Go

قطاع غزة دون أن يسفر ذلك عن وقوع إصابات وكان قد سمع دوي انفجارين قويين عند نحو الساعة السادسة صباحا بالتوقيت المحلي اليوم المستوطنة المذكورة وأشارت الأنباء إلى أن عدة انفجارات سمعت بوضوح في مدينة غزة التي تبعد بضعة كيلومترات عن المستوطنات. وقد قامت القوات الإسرائيلية على اثر ذلك بقصف المناطق المجاورة للمستوطنة بقذائف الدبابات كما جرى تبادل لإطلاق النار مع مسلحين فلسطينيين. هذا وقد أعلن مصدر في الشرطة الاسرائيلية اليوم الجمعة أن القوات الإسرائيلية ستبقى أو ستبقي حالة الاستنفار القصوى خشية وقوع هجمات فلسطينية حتى نهاية أسبوع الفصح اليهودي في الرابع عشر من الشهر الجاري. من جهة ثانية أسفر القصف الصاروخي الذي نفذته المروحيات الإسرائيلية ويستهدف أهدافا فلسطينية في قطاع غزة عن إصابة ستة فلسطينيين.	Khadijah Bin-Qinnah
الغارة الإسرائيلية على غزة في إطار التصعيد الإسرائيلي والحملة المستمر ضد الفلسطينيين سواء بالقصف ومحاولاته فياض. ولم تكف توحدت قصف ليلي حتى سرعة الدبابة الإسرائيلية إلى قصف المناطق المحيطة بغزة بعد تعرض مستوطنة تنساريم للقصف بقذائف الهاون. الغارة الاسرائيلية اسفرت عن اصابة ستة فلسطينيين وخلفت دمارا واسعا المواقع التي استهدفتها ومن مقر رئاسة قوات الشرطة الفلسطينية في بيت لاهيا الذي دمر تماما إلى جانب أربعة مراكز لقوات الأمن الفلسطيني قرب مخيم جباليا لحقت بها اضرار جسيمة. في غضون ذلك عززت إسرائيل إجراءاتها الأمنية في مدينة القدس وخاصة في محيط الحرم القدسي الشريف تحسبا لتصاعد المواجهات بعد صلاة الجمعة كما عزز الجيش الإسرائيلي من نقاط التفتيش على مداخل مدينتي القدس وبيت. ويأتي الإجراء الإسرائيلي عقب تهديدات لحركة الجهاد الإسلامي بشن عمليات إسرائيل انتقاما لاغتيال أحد قادة ياتي كينيين. وقالت الشرطة الإسرائيلية إنه سيتم الإبقاء على حالة الاستنفار القصوى خشية وقوع هجمات حتى نهاية اسبوع الفصح اليهودي الذي ينتهي في الرابع عشر من الشهر الجاري.	male 2
هذا وقد اقترح وزير الأمن الداخلي الإسرائيلي عوزي لاندو اليوم الجمعة تكثيف الغارات على الفلسطينيين بمن فيهم الشخصيات في السلطة الفلسطينية إذا كان لهم دور في الهجمات التي تتعرض لها أهداف إسرائيلية. وجاء تصريح الوزير المتشدد إثر تصفية أجهزة الأمن الإسرائيلية أحد أبرز قادة الجناح العسكري لحركة الجهاد الإسلامي في فلسطين وذلك في جنين. وبعد سلسلة حوادث تبادل لإطلاق نار وقصف خلال الليل. وتأتي هذه التطورات بعد ساعات من قيام جنود	Khadijah Bin-Qinnah

Figure 4: STT output with automatic sentence boundary detection

Figure 4 shows the output of our Arabic Audio Indexer System for an Al-Jazeera TV episode, automatically annotated with sentence boundaries by the *Combined* SBD system. The periods denoting the hypothesized sentence boundaries are followed by double spaces to show the enhancement in readability of the Arabic STT transcripts with automatic sentence boundary detection.

5. Conclusions

In this paper, we have described an automatic language-independent system to detect sentence boundaries in Arabic speech recognition transcripts. The output of our Arabic Audio Indexer system with automatic sentences shows the value added by our sentence boundary detection system to the speech transcripts. It is very difficult for us to compare to existing research, which has been focused on English. We plan to apply the same approach, as described in this paper, to the English BN, and CTS domains for effective comparisons. We also plan to use large amounts of text transcriptions to build statistical language models that improve the performance of the linguistic system. We believe that the performance of our *Combined* SBD system, at approximately 50% DER, is indicative of a useful sentence-based speech transcript.

6. Acknowledgement

This work was supported in part by The Office of Advanced Information Technology under Contract 1999-S0189000-000 and by DARPA under Contract N66001-00-C-8008.

7. References

[1] Heidi Christensen, Yoshihiko Gotoh, and Steve Renals, “Punctuation Annotation using Statistical Prosody Model”, Proceedings of Eurospeech, Aalborg, Denmark, 2001.

[2] Jim Huang, and Geoffrey Zweig, “Maximum Entropy Model for Punctuation Annotation from Speech”, Proceedings of ICSLP, pp 917-920, Denver, Colorado, 2002.

[3] Ji-Hwan Kim, and P. C. Woodland, “The Use of Prosody in a Combined System for Punctuation Generation and Speech Recognition”, Proc. Eurospeech, 2001.

[4] E. Shriberg, et al., “Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?”, Language and Speech, pp 439-487, 1998.

[5] E. Shriberg, A. Stolcke, D. Hakkani-Tur, G. Tur, “Prosody-Based Automatic Segmentation of Speech into Sentences and Topics”, Speech Communications, Special Issue on Accessing Information in Spoken Audio, 2000.

[6] Dan Bikel, Scott Miller, Richard Schwartz, Ralph Weischedel, “Nymble: A High-Performance Learning Namefinder,” Proceedings of the Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics, pp. 194-201, 1997.

[7] J. Billa, M. Noamany, A. Srivastava, D. Liu, R. Stone, J. Xu, J. Makhoul, F. Kubala, “Audio Indexing of Arabic Broadcast News”, Proceedings of ICASSP, Orlando, Florida, 2002.

[8] John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel, “Performance Measures for Information Extraction”, Proc. Of DARPA Broadcast News Workshop, pp 249-252, Herndon, VA, February 1999.