4th ASA/ASJ Joint Meeting Lay Language Papers

More Than Words Can Say:
Using prosody to find sentence boundaries in speech

Yang Liu
University of Texas at Dallas
Richardson, Texas, 75080, USA
yangl@hlt.utdallas.edu

Elizabeth Shriberg
SRI International and International Computer Science Institute
Menlo Park, California, 94025, USA
ees@speech.sri.com

Popular version of paper 1aSC2.
Presented on 28th of November, morning, 2006.
4th ASA/ASJ Joint Meeting, Honolulu, HI

Automatic speech recognition has improved dramatically over the past few decades. The goal of such systems is typically only to output a simple stream of words. Humans, however, use information beyond words alone to convey and comprehend messages. Such information includes intonation (variations in pitch), rhythm and timing of speech sounds, stress and loudness patterns, and even variation in voice quality. Together, these cues are referred to as *prosodic* information. For example, prosodic information is crucial to determining meaning in the following examples:

(a) John likes Mary. John likes Mary?
(b) JOHN likes Mary. John likes MARY.
(c) No. Dogs are allowed. No dogs are allowed.

Prosody is also used by speakers to convey and detect emotional states, level of interest, and so on. The use of prosody for all of these functions is integral to speech across languages, and not limited to the otherwise ambiguous examples shown above. One way to see this clearly is in synthesized speech, which despite great strides in intelligibility, still often sounds unnatural.

In this work, we focus on the function in example (c), that is, we examine how to harness prosody to help automatically identify sentence boundaries in speech. Speech recognition output generally does not contain any punctuation marks. Although a speaker using a dictation system may train him- or herself to explicitly utter “comma”, “semi-colon”, “period”, this is a particularly unnatural way of speaking. People talking with each other do not do this at all. Instead, they convey punctuation *implicitly*, via prosodic and syntactic information.

To identify sentence boundaries in speech, we use both words and prosodic information. Words and syntactic information are no doubt very important to signal the end of a sentence. However, as example (c) above shows, words themselves might be ambiguous and, in these cases, prosody may provide additional valuable information to help in understanding speech. In this paper, we focus on the role of prosody for sentence boundary detection.

Figure 1 shows the speech for two examples: “No. Dogs are here.” (on the left) and “No dogs are here.” (on the right). The upper part of the two examples displays the variation in sound pressure over time; the lower part shows the pitch contour. Typically, there is a pitch drop (“declination”) at the end of a sentence. We can see from the figure that there is a pause after “No.” on the left graph, whereas there is not such a long pause in the second example between “no” and “dogs”. In addition, it can be seen that there is a rising pitch contour for “no” in the second example.

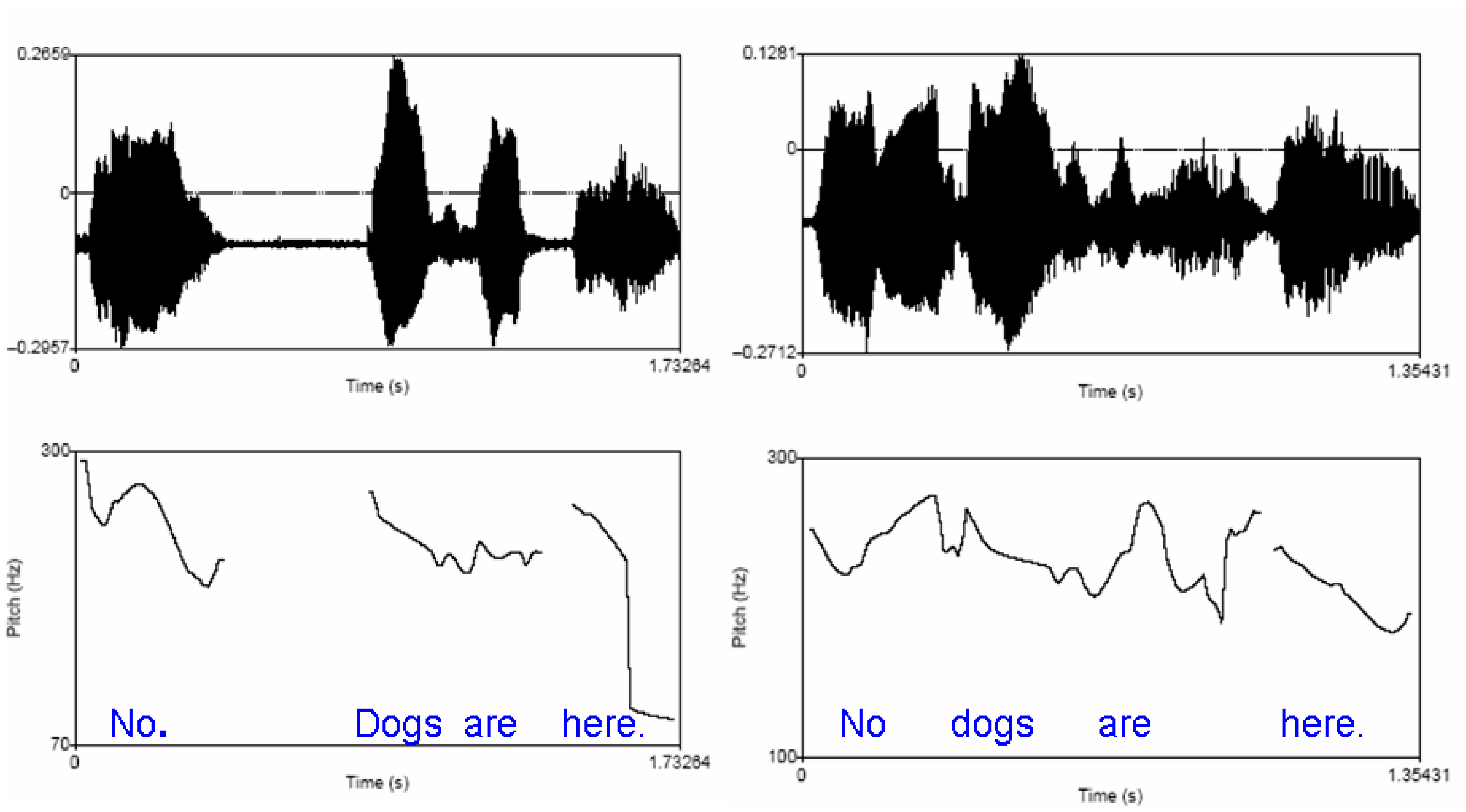

Figure 1: Using prosody to help identify sentence boundaries and disambiguate meanings.

We extract these prosodic cues automatically from the audio data, using the word, syllable, and phone boundaries from either a human transcription force-aligned by a speech recognizer, or from a speech recognizer itself. Since sentence boundaries only occur between words, our tokens are inter-word boundaries. Each inter-word boundary is associated with a vector of prosodic features (e.g., pause length [if any], relative duration of phones preceding the boundary, pitch contours before and after the boundary). We may use more than 100 features in our experiments, normalized for inherent variation in speech sounds and also for the particular talker [1].

We then use machine learning techniques (e.g., a classifier) to learn which features (and which combinations of features) are best for discriminating sentence-boundary tokens from intra-sentence tokens. During testing, the classifier hypothesizes whether or not there is a sentence boundary, based on the prosodic cues. In fact, the classifier can provide not just a binary output, but rather a score that can then be thresholded depending on one’s tolerance for false alarms versus false rejections.

We have performed experiments on both conversational speech and broadcast news speech. There is a clear speaking style difference between these two corpora. For example, sentences in broadcast news are generally longer and more syntactically complex than those in conversational speech. Conversational speech contains high rates of discourse-related elements, such as “uh-huh” and “you know”, as well as disfluencies. In addition, while speakers in natural conversation display significant individual differences in prosodic style, news anchors in read broadcasts tend to adopt prosodic patterns that are similar to those of other anchors.

Table 1 shows the classification error rate of sentence boundary detection using human transcriptions for the two different styles of speech [2]. We present results when using prosodic information alone, word information only, and using the combination of the two knowledge sources. The results show that prosody provides additional information beyond words and helps better identify sentence boundaries and disambiguate meanings.

	Broadcast News	Conversations
Using prosody only	5.4%	8%
Using words only	5.4%	5.5%
Using words + prosody	4%	4.4%

Table 1: Classification error rates for sentence boundary detection on broadcast news and conversational speech, using prosody and words alone, and their combination.

In addition to finding sentence boundaries, prosody is useful in distinguishing whether an utterance is a statement or a question, as in example (a). In spontaneous conversation, it is exceedingly common in English to ask a question using declarative syntax (“It’s raining?”) rather than interrogative syntax (“Is it raining?”).   Knowing whether a declarative utterance like “It’s raining” is a statement or a question is quite important, since the listener’s response differs depending on the sentence modality.   Figure 2 shows examples of the word sequence “I’m sorry”. In the first example (on the left in Figure 2), the question intonation conveys that the speaker is asking for clarification of something previously said, whereas in the second case (on the right in Figure 2) the speaker is apologizing. The difference can be clearly seen in the intonation contour as well as the envelope of the sound pressure pattern on “sorry”.

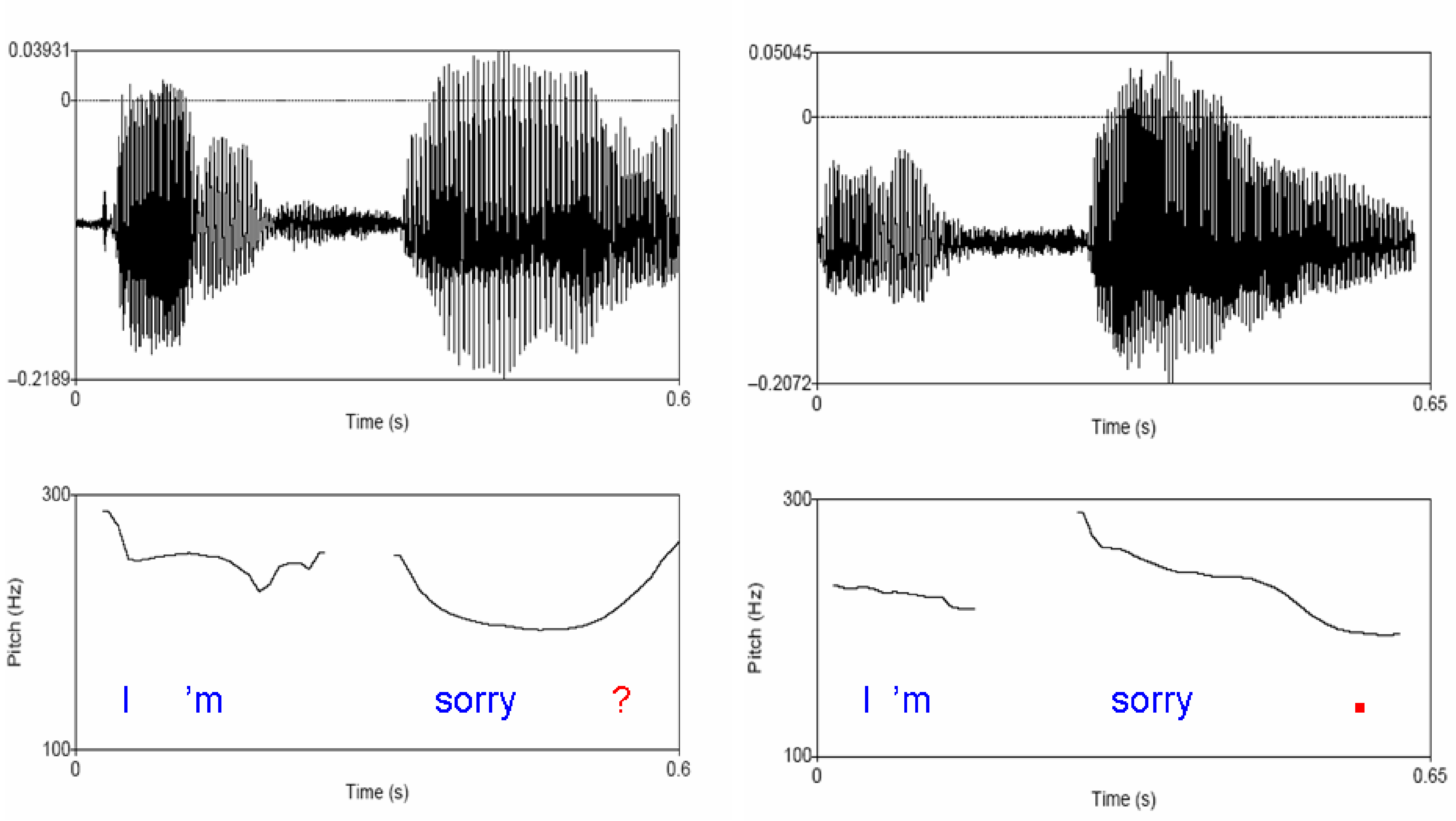

Figure 2: "I'm sorry?" with a rising tone and “I’m sorry.” with a falling tone.

An interesting new area in automatic prosody modeling is how to cope with individual variation. In a recent study on sentence boundary detection in naturally occurring multiparty meetings [3], we observed considerable speaker differences in how speakers conveyed sentences prosodically. For example, some speakers made use of long pauses between sentences, while others tended to use shorter pauses but stronger intonation cues. We also noticed that there seemed to be a difference between native and nonnative speakers when conversing in English, which is interesting because prosody is often difficult for second-language learners to master.

Acknowledgments:

We thank Andreas Stolcke, Jachym Kolar, Dustin Hillard, Mari Ostendorf, and Mary Harper for collaborating on this work, which was supported by DARPA under contracts No. MDA972-02-1-0013, HR0011-06-C-0023, and NBCHD030010, and AMI FP6-506811.

References

[1] Yang Liu, Nitesh Chawla, Mary Harper, Elizabeth Shriberg, and Andreas Stolcke, "A Study in Machine Learning from Imbalanced Data for Sentence Boundary Detection in Speech", *Computer Speech and Language*, v20(4), pp468-494, 2006.

[2] Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper, “Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies”, *IEEE Transactions on Audio, Speech, and Language Processing*, v14(5), pp1526-1540, 2006.

[3] Jachym Kolar, Elizabeth Shriberg, and Yang Liu, "On Speaker-Specific Prosodic Models for Automatic Dialog Act Segmentation of Multi-Party Meetings", *Proceedings of Interspeech*, Pittsburgh, 2006.