Spontaneous Speech: How People Really Talk and Why Engineers Should Care

Elizabeth Shriberg

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA 94025, USA
International Computer Science Institute, Berkeley, CA 94704, USA
ees@speech.sri.com

Abstract

Spontaneous conversation is optimized for human-human communication, but differs in some important ways from the types of speech for which human language technology is often developed. This overview describes four fundamental properties of spontaneous speech that present challenges for spoken language applications because they violate assumptions often applied in automatic processing technology.

1. Introduction

Most of the speech we produce and comprehend each day is spontaneous. This "speech in the wild" requires no special training, is remarkably efficient, imposes minimal cognitive load, and carries a wealth of information at multiple levels. Spontaneous conversation is optimized for human-human communication, but differs in some important ways from the types of speech for which human language technology is often developed [33]. This overview describes four fundamental properties of spontaneous speech that present problems for spoken language applications because they violate assumptions often applied in automatic processing technology. While not meant to be exhaustive, the four properties are chosen because they are either pervasive or important to spoken language applications when they do occur, and because they represent difficult and interesting challenges to speech science and engineering. Together they cover a range of phenomena—from lower-level linguistic structure to higher-level phenomena.

10.21437/Interspeech.2005-3

2. Four Fundamental Challenges

2.1. Recovering hidden punctuation

In many formal written languages, punctuation is rendered explicitly. But spoken language is a stream of words, with no overt lexical marking of the punctuation itself. Instead, phrasing is conveyed through other means, including prosody [16]. Sentence boundaries and other types of punctuation are useful for many types of automatic downstream processing (including parsing, information extraction, dialog act modeling, summarization, and translation), as well as for human readability [40, 47, 18, 24, 22, 33, 21, 32]. These methods are typically trained on text data, which contains punctuation. Modeling sentence-level punctuation can also improve speech recognition performance itself [44, 24].

Historically, speech recognition researchers have built language models based on sentences as found in text and then tried to acoustically segment the speech into sentence-like units. This is typically done by chopping at longer pauses and speaker changes. Pauses are relatively easy to detect and minimize the risk of fragmenting words in the process. Speaker changes are also often available, especially if speakers are recorded on different channels. For some applications, if a speaker produces one sentence at a time (for example, to a dialog system) there is typically little problem. But for processing of natural conversation, finding the sentence boundaries by machine is a challenge. Pauses are neither necessary nor sufficient indicators of sentence boundaries. People often string together sentences without pauses. And conversely, people pause (as during hesitations or disfluencies) at locations other than sentence boundaries.

Computational models for finding sentence boundaries in speech typically involve a combination of N-gram language models (over words and boundary labels) and prosodic classifiers [46, 40]. Knowledge sources are often combined using an HMM framework. More recently, other model types have been used successfully, such as maximum entropy models and conditional random fields [31]. Prosodic models have used probabilistic classifiers such as decision trees or neural networks, and can be improved by sampling and ensemble techniques [29]. Additional approaches and features are described in [45]. While initial research used hand-transcribed words as input, more recent work has studied the problems arising from imperfect recognizer hypotheses. In particular, work on strategies for using multiple recognition hypotheses appears promising [14].

The tasks described above typically apply to offline processing of speech intended for human listeners. But there is an important application of punctuation detection for online human-computer dialog systems: "endpointing", or determining when a speaker has completed an utterance to a system. The endpointing task has some relationship to turn-taking (a topic discussed further in Section 2.3), but is described here because it requires the same basic disambiguation between hesitation and grammatical pauses that is involved in offline punctuation work. Most current endpointers wait for a pause that is longer than some predetermined duration threshold. But this assumption that a pause signals the end of a speaker's turn is violated when the speaker pauses while hesitatating. The challenge in endpointing, as compared with offline sentence segmentation, is that endpointing is an online task and must be accomplished using information only *before* the potential endpoint under consideration.

An approach to reducing cutoffs during hesitation and also for speeding responses after true utterance ends is described in [9]. The approach models prosodic features occurring before pauses. Not only do these features significantly reduce premature cutoffs during hesitations, they also provide a remarkable reduction of up to 81% for the "average speaker waiting time" for true utterance ends. Prosodic features such as intonation add confidence to true ends even before a speaker pauses, thus

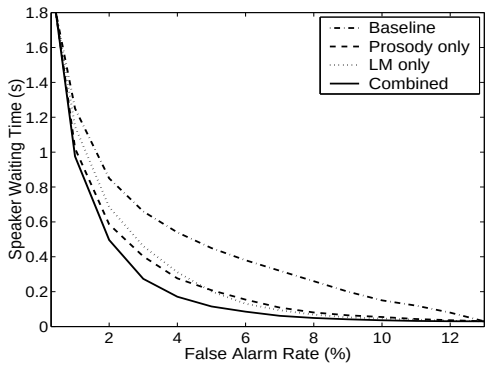

Figure 1: Speaker Waiting Time (SWT) vs. False Alarm Rate (FAR) for four endpointing systems, from [9]. Systems, ordered from worst to best performance, are: baseline system (pause-only), language-model system, prosody system, combined prosody + language model system.

shortening the required pause threshold for a given false alarm rate and speeding the overall interaction.

2.2. Coping with disfluencies

Disfluencies such as filled pauses, repetitions, repairs, and false starts are frequent in natural conversation. Across many corpora and languages, disfluencies occur at rates higher than every 20 words, and can affect up to one third of utterances [38]. Although disfluencies were once viewed as "errors", a growing literature on the topic has come to appreciate them as an integral part of natural language and conversation. They reflect both cognitive aspects of language production and the management of interaction [26, 38, 5].

Disfluencies are important to model in speech technology because they cause problems for higher level natural language processing, such as parsing, translation, information extraction, and summarization; they also degrade transcript readabililty for humans [47, 11, 3, 21]. In some contexts, modeling disfluencies can also improve speech recognition performance [44, 28]. Most of the approaches to the natural language processing problems are based on training using large amounts of written text, which typically does not contain disfluencies. One strategy is thus to detect and then remove (or "clean up") the disfluencies prior to further processing.

Detecting filled pauses can be nearly trivial for words such as "um" in English, if recognition output is correct, but is difficult for cases like "uh" (homophonous with "a"), or for words that can function either as a filler or nonfiller. A common approach to edit detection is to first detect the interruption point of a disfluency using lexical and prosodic cues, and to then use other cues to determine the extent of the excision region. Early work [15] assumed that the interruption point is known and focused on identifying the reparandum and repair regions. Finding the interruption point, however, is a nontrivial problem in itself. Some words, such as filled pauses and editing terms, are direct evidence, but in general these cues are ambiguous and more powerful machine learning techniques must be used.

In [13], a statistical model is developed for the correspondence between words in the reparandum and repair regions for disfluency detection. More recently, a source-channel model has been used, together with a tree adjoining grammar [20] or a statistical model to represent the possibility of a word being disfluent [17]. Transformation-based learning is another data-driven learning paradigm that has been applied to the problem [41]. Finally, the same basic techniques employed for sentence boundary detection have also been applied to process disfluencies [30].

Table 1: Percentage of overlapped speech units in different corpora, for words (in plainface) and "spurts" (stretches of speech that do not contain any pauses greater than 500 msec.; in italics). "MR" meetings are discussions; "ROB" meetings are directed by a speaker who produces 60% of all words, and tend toward individual reporting. "CH" = CallHome (friends and family, real calls); "SWB" = Switchboard (strangers paid to discuss prescribed topics).

	Meetings		Phone convs.	
Backchannels	MR	ROB	CH	SWB
Included				
words	17.0	8.8	11.7	12.0
spurts	*54.4*	*31.4*	*53.0*	*54.4*
Excluded				
words	14.1	5.6	7.9	7.8
spurts	*46.4*	*21.0*	*38.8*	*38.9*

2.3. Allowing for realistic turn-taking

Spontaneous speech has another dimension of difficulty for automatic processing when more than one speaker is involved. As described in classic work on conversation analysis [36], turn-taking involves intricate timing. In particular, speakers do not alternate sequentially in their contributions as often suggested by the written rendition of dialog. Rather, listeners project the end of a current speaker's turn using syntax, semantics, pragmatics, and prosody, and often begin speaking before the current talker is finished [36, 16, 4, 27].

Overlap is frequent not only in corpora like multiparty meetings, but also in two-party phone conversations, as shown in Table 1 [39]. Overlap rates are high even when backchannels (such as "uh-huh") are removed from consideration. What is important is that overlap in two-party phone conversations is not necessarily lower than in meetings. It is also interesting that familiarity with the other talker does not appear to affect the rate of overlap, since there is no difference in overall rates between the CallHome and Switchboard conversations.

Modeling realistic turn-taking has a fascinating application to dialog system design; some researchers are developing conversational systems that can mimic human turn-taking by providing backchannel responses, e.g., [10]. Overlap in turn-taking also introduces several problems for many current offline automatic speech processing tasks. An obvious difficulty is the acoustic modeling of simultaneous speakers on a single recording channel. Relatively little work has focused on this problem to date, although source separation and auditory scene analysis techniques may ultimately lead to solutions. In this paper we focus on the impact of overlap for higher-level phenomena.

One area is language modeling. Conversational speech recognition work has largely focused on the case in which each speaker is recorded on a separate channel, and the channel is modeled separately as an independent stream of words. However, recent work has shown that conditioning word prediction across speakers leads to improvements [19]. The situation becomes complicated when, because of overlap, a given word is immediately preceded by words from both the same and another speaker. Further complications arise in multiparty meetings, when several other speakers could provide the preceding words to condition on. Lack of strictly sequential turns also causes problems for the automatic classification of dialog acts. Work has shown that in telephone conversations, the dialog context (the preceding and following utterances, and who said them)

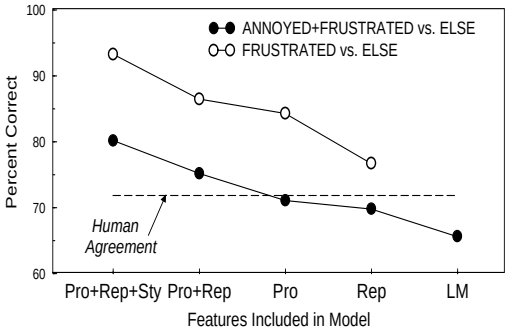

Figure 2: Annoyance and frustration detection results for a telephone dialog system, from [1]. Pro = prosody features (automatic), Sty = hyperarticulation-related features (hand-coded), Rep = repetition/correction feature (hand-coded). Human agreement is on the more difficult task (Annoyed+Frustrated vs. Else).

can be an important source of information in disambiguating dialog acts [43]. The models used for this purpose are typically sequential (e.g., HMMs), and it is not clear how to best generalize them to the case of overlapping speech.

Even *evaluating* speech recognition in the face of simultaneous speech is not straightforward. NIST meeting recognition evaluations in the United States have until recently simply excluded regions of overlapping speech from word error scoring. The method used is computationally demanding, since a single hypothesized stream of words has to be aligned to multiple reference streams [12].

2.4. Hearing more than words

A fourth challenge area is to "hear" a speaker's emotion or state of being, through speech. Modeling emotion and user state is particularly important for certain dialog system applications. For example, in an automated assistance application, one would want to transfer an angry user to a human. Detecting affect through speech obviously requires more than just words. Despite a growing literature in both linguistics and applied fields, this area remains a challenge both because it is such an inherently difficult task, and because it is hard to obtain natural emotion data [25, 1, 7, 2, 37, 8, 35, 42, 6]. Most data comes from acted speech, which is usually not a good predictor of natural emotions, and any natural data that is collected has to be labeled for emotion, which is difficult for humans to do [1, 42, 6]. Relatively few studies examine the detection of non-acted emotions in a realistic setting, using an automatic dialog system and automatic speech recognition output.

One exception is a large-scale study of natural emotion in interaction with a telephone dialog system [1]. The study investigated both human-human and human-computer agreement on detection of annoyance and frustration. Automatic detection was explored for both true and recognized words. In addition to using features based on lexical and prosodic information, experiments examined the effect of two other types of features hypothesized to be important in this context: hyperarticulation (often used after recognition errors) and repeated attempts at getting the same information (obviously correlated with frustration). These last two feature types were hand-coded, since the goal was to find out how the features interacted with automatically extracted lexical and prosodic features. Results are shown in Figure 2.

A number of points can be noted. First, performance is better for the more extreme emotions. This is not surprising, but given the much smaller amount of training data for this task, the effect is quite large. Second, machine performance against a second-pass consensus labeling by multiple human labelers is similar to the agreement obtained over individual labelers. Machine accuracy is about equal to human agreement when only prosodic features are used, and it is even better than human agreement when the additional features are included. Finally, in this type of data, word information alone (other than the occasional expletive) is a rather poor predictor of emotional state.

3. Future Directions

All four phenomena discussed would benefit from improved basic features, as well as from better methods for integrating knowledge from different feature types. For both lexical and prosodic features, another challenge is robustness to speech recognition errors. To optimize downstream performance in an application (for example, to best use sentence boundaries for parsing, or emotion recognition for a tutoring system), it will be important to preserve multiple hypotheses or use soft decisions. And for work in multimodal recognition, improvements in all four areas could come from integration of spoken and visual information.

The first three phenomena could also benefit from the incorporation of longer-range information than is typically currently modeled. This applies to both language modeling (where largely only N-grams are used) and prosodic modeling (where longer-range features should offer more than local features, but would also increase the complexity of the search). Another challenge for these phenomena is to move from the current tendency to treat target events as independent, to techniques that can model dependencies both within event streams and across related events.

Finally, a promising area for further work is speaker-dependent modeling. Recent work in speaker recognition, albeit a different task, provides evidence that individuals differ not only in frame-level acoustics, but also in habitual lexical and prosodic patterns [34, 23]. Since some of these patterns are related to the types of features used in modeling the four phenomena discussed here, it is likely that some sort of style-based adaptation can improve performance in these areas. To take an example of speaker differences: people vary not only quantitatively but also qualitatively in disfluency production [38, 17]. A distinction between "repeaters" (people who tend to produce more repetitions than deletions or false starts) and "deleters" (people who show the opposite pattern) was found in [38]. Deleters were furthermore faster speakers than repeaters, suggesting that the groups may employ different speaking strategies (with deleters sometimes getting ahead of their thoughts and having to start anew). Such differences are relevant to technology applications because repetitions are much easier to handle in automatic processing than are deletions.

4. Conclusions

This overview described four challenge areas for the automatic modeling of spontaneous speech. In each area, speakers convey useful information on multiple levels that is often not modeled in current speech technology. Greater attention to these challenges, as well as increased scientific understanding of natural speaking behavior, should offer long-term benefits for the development of intelligent spoken language applications.

5. Acknowledgments

I thank the Interspeech 2005 organizers for their invitation, and Yang Liu and Andreas Stolcke for many helpful contributions. This work was funded by NSF IRI-9619921 and IIS-012396, and by DARPA contracts MDA97202C0038 and NBCHD030010. The views herein are those of the author and do not reflect the views of the funding agencies. Distribution is unlimited.

6. References

[1] J. Ang et al. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. ICSLP*, 2002.

[2] A. Batliner et al. How to find trouble in communication. *Speech Communication*, 40, 2003.

[3] C. Boulis et al. The role of disfluencies in topic classification of human-human conversations. In *AAAI Workshop on Spoken Language Understanding*, 2005.

[4] J. Caspers. Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics*, 31(2):251–276, 2002.

[5] H. Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.

[6] L. Devillers et al. Challenges in real-life emotion annotation and machine learning based detection. *Journal of Neural Networks*, 18(4), 2005.

[7] E. Douglas-Cowie et al. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40:33–60, 2003.

[8] R. Fernandez and R. Picard. Classical and novel discriminant features for affect recognition from speech. In *Proc. Interspeech*, 2005.

[9] L. Ferrer et al. A prosody-based approach to end-of-utterance detection that does not require speech recognition. In *Proc. ICASSP*, 2003.

[10] S. Fujie et al. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialog system. In *Proc. Interspeech*, 2005.

[11] S. Furui et al. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Trans. Speech and Audio Process.*, 12(4):401–408, 2004.

[12] J. S. Garofolo et al. The Rich Transcription 2004 Spring meeting recognition evaluation. In *Proc. NIST ICASSP 2004 Meeting Recognition Workshop*, 2004.

[13] P. A. Heeman and J. F. Allen. Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialog. *Computational Linguistics*, 25(4):527–571, 1999.

[14] D. Hillard et al. Improving automatic sentence boundary detection with confusion networks. In *Proc. HLT-NAACL*, 2004.

[15] D. Hindle. Deterministic parsing of syntactic non-fluencies. In *Proc. ACL*, 1983.

[16] J. Hirschberg. Communication and prosody. *Speech Communication*, 36, 2002.

[17] M. Honal and T. Schultz. Automatic disfluency removal on recognized spontaneous speech—rapid adaptation to speaker-dependent disfluencies. In *Proc. ICASSP*, 2005.

[18] J. Huang and G. Zweig. Maximum entropy model for punctuation annotation from speech. In *Proc. ICSLP*, 2002.

[19] G. Ji and J. Bilmes. Multi-speaker language modeling. In *Proc. HLT-NAACL*, 2004.

[20] M. Johnson and E. Charniak. A TAG-based noisy channel model of speech repairs. In *Proc. ACL*, 2004.

[21] D. Jones et al. Measuring human readability of machine generated text: Three case studies in speech recognition and machine translation. In *Proc. ICASSP*, 2005.

[22] J. Kahn et al. Parsing conversational speech using enhanced segmentation. In *Proc. HLT-NAACL*, 2004.

[23] S. S. Kajarekar et al. SRI's 2004 NIST speaker recognition evaluation system. In *Proc. ICASSP*, 2005.

[24] J.-H. Kim and P. C. Woodland. A combined punctuation generation and speech recognition system and its performance enhancement using prosody. *Computer Speech and Language*, 41(4):563–577, 2003.

[25] C. M. Lee et al. Combining acoustic and language information for emotion recognition. In *Proc. ICSLP*, 2002.

[26] W. J. M. Levelt. Monitoring and self-repair in speech. *Cognition*, 14:41–104, 1983.

[27] A. J. Liddicoat. The projectability of turn constructional units and the role of prediction in listening. *Discourse Studies*, 6(4):449–469, 2004.

[28] C. K. Lin and L. S. Lee. Improved spontaneous Mandarin speech recognition by disfluency interruption point (IP) detection using prosodic features. In *Proc. Interspeech*, 2005.

[29] Y. Liu et al. Using machine learning to cope with imbalanced classes in natural speech: Evidence from sentence boundary and disfluency detection. In *Proc. ICSLP*, 2004.

[30] Y. Liu et al. Comparing HMM, maximum entropy, and conditional random fields for disfluency detection. In *Proc. Interspeech*, 2005.

[31] Y. Liu et al. Comparing and combining generative and posterior probability models: Some advances in sentence boundary detection in speech. In *Proc. EMNLP*, 2004.

[32] J. Makhoul et al. The effects of speech recognition and punctuation on information extraction performance. In *Proc. Interspeech*, 2005.

[33] M. Ostendorf et al. Human language technology: Opportunities and challenges. In *Proc. ICASSP*, 2005.

[34] D. Reynolds et al. The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition. In *Proc. ICASSP*, 2003.

[35] M. Rotaru and D. Litman. Using word-level pitch features to better predict student emotions during spoken tutoring dialogues. In *Proc. Interspeech*, 2005.

[36] H. Sacks et al. A simplest semantics for the organization of turn-taking in conversation. *Language*, 50(4):696–735, 1974.

[37] B. Schuller et al. Speaker-independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Proc. Interspeech*, 2005.

[38] E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, U.C. Berkeley, 1994.

[39] E. Shriberg et al. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Proc. EUROSPEECH*, 2001.

[40] E. Shriberg et al. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154, 2000.

[41] M. Snover et al. A lexically-driven algorithm for disfluency detection. In *Proc. HLT-NAACL*, 2004.

[42] S. Steidl et al. Of all things the measure is man: Automatic classification of emotions and inter-labeler consistency. In *Proc. ICASSP*, 2005.

[43] A. Stolcke et al. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, 2000.

[44] A. Stolcke et al. Modeling the prosody of hidden events for improved word recognition. In *Proc. EUROSPEECH*, 1999.

[45] D. Wang and S. Narayanan. A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues. In *Proc. ICASSP*, 2004.

[46] V. Warnke et al. Integrated dialog act segmentation and classification using prosodic features and language models. In *Proc. EUROSPEECH*, 1997.

[47] K. Zechner. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4), 2002.