Enhanced End-of-Turn Detection for Speech to a Personal Assistant

Harish Arsikere[1] **Elizabeth Shriberg[2]** **Umut Ozertem[3]**

[1]*Data Analytics Lab, Xerox Research Center India (XRCI), Bangalore, Karnataka, India*
[2]*Speech Technology and Research Laboratory, SRI International, Menlo Park, California, USA*
[3]*Microsoft Corporation, Sunnyvale, California, USA*
{harish.arsikere@xerox.com; elizabeth.shriberg@sri.com; umut.ozertem@microsoft.com}

Abstract

Speech to personal assistants (e.g., reminders, calendar entries, messaging, voice search) is often uttered under cognitive load, causing nonfinal pausing that can result in premature recognition cut-offs. Prior research suggests that prepausal features can discriminate final from nonfinal pauses, but it does not reveal how speakers would behave if given longer to pause. To this end, we collected and compared two elicitation corpora differing in naturalness and task complexity. The *Template Corpus* (4409 nonfinal pauses) uses keyword-based prompts; the *Freeform Corpus* (8061 nonfinal pauses) elicits open-ended speech. While nonfinal pauses are longer and twice as frequent in the Freeform data, prepausal feature modelling is roughly equally effective in both corpora. At a response latency of 100 ms, prepausal features modelled by an SVM reduced cut-off rates from 100% to 20% for both corpora. Results have implications for enhancing turn-taking efficiency and naturalness in personal-assistant technology.

Introduction

Conventional spoken dialog systems endpoint turns using a fixed-duration nonspeech threshold, typically on the order of 500ms. A problem with such systems is the high rate of premature user cut-offs caused by medial or nonfinal pauses (especially the longer ones), which occur quite frequently in natural speech directed to personal assistants (Arsikere et al., 2014). Cut-off rates could be particularly high for mobile applications in which users are multi-tasking and under cognitive load. Simply increasing the nonspeech threshold is not a viable solution, since it will lead to additional latency at actual ends of turns.

Previous research on dialog systems has explored end-of-turn detection using acoustic, prosodic and lexical cues (Hariharan et al., 2001; Ferrer et al., 2002, Ferrer et al., 2003; Raux and Eskenazi, 2012; Arsikere et al., 2014). Related work has studied the detection of hesitations and disfluencies in human-human dialog (Liu et al., 2006; O'Shaughnessy, 1992; Audhkhasi et al., 2009).

In this study we ask how well premature cut-offs can be prevented at low system latency, by using acoustic cues before pause onset. We are particularly interested in how speakers would behave *if they did not need to worry about being cut off during pauses.* Using an existing dialog-system collection is thus not ideal, since if it involved a standard cut-off threshold, any speech after longer pauses would be absent from the collection. Note that simply computing cut-off rates in such data does not convey the cost of speakers taking longer to start their turns, or chunking content into multiple smaller turns.

To this end, we collected two corpora of elicited speech using tasks differing in naturalness and cognitive load. The Template Corpus was first studied in Arsikere et al. (2014); the Freeform Corpus is introduced in this paper. Features and models are held constant over the two corpora, and they are compared with regard to pausing behaviors, model performance (accuracy of endpoint detection) and feature usage. The present work is novel in several ways compared to previous efforts. It allows and deals with long nonfinal pauses unlike Hariharan et al. (2001); it does not rely on a speech recognizer or session information for computing acoustic features unlike the studies by Ferrer et al.; and, most importantly, whereas Arsikere et al. (2014) studied speech elicited using a "fill-in-the-blank" task, the current paper uses a more naturalistic, open-ended elicitation.

Method

Data collection

The subjects were adult native speakers of American English. A tool displayed a set of prompts one at a time in random order on a screen. Table 1 shows sample prompts for the Template and Freeform corpora. We adapted a commercial message-dictation back-end by increasing the

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved. The corpora used were collected while the first author was interning at Microsoft. We thank A. Fidler, P. Parthasarathy, M. Petruck, and K. Larsen, for corpus assistance.

nonspeech threshold tenfold so that users would not be cut off, but would see actual recognition output (not error free). Subjects had a fixed amount of time after the prompt in which to start speaking (3 seconds for Template, 5 seconds for Freeform, based on a pilot study of appropriate delays to allow comprehension but not full pre-planning). Subjects were asked to speak as if they were talking to a personal assistant, using content from their personal lives that would make sense given the prompt. They then saw the recognition hypothesis, and when ready moved to the next prompt. Responses were recorded in a quiet environment using a close-talking microphone, at a sampling rate of 16 kHz and a resolution of 16 bits/sample.

Table 1: Sample prompts from each corpus.

Template Corpus	Answer email from <PERSON> about <TOPIC>.
Freeform Corpus	GOAL: *Create calendar entry.* SCENARIO: An outing with friends (place, date, time).

End-of-turn decisions were made at all pauses longer than 100 ms; this is roughly the shortest duration that is longer than most stop bursts. For convenience, the nonfinal pauses were obtained using a message-dictation back-end. This may overestimate the alignment accuracy provided by a speech/nonspeech detector, but the assumption is that the general pattern of results should hold. Table 2 summarizes the statistics for the two corpora used. As shown, while the Freeform corpus has fewer speakers and utterances, it has a much higher rate of pausing and longer pause durations.

The prompt categories were reminders, calendar entries, messaging (SMS/email), and voice search – common use cases for a personal assistant. The template data tended to follow the prompts, e.g., "Answer email from Emily about cookies", "Answer email from Steven about share prices". Freeform utterances, on the other hand, were more variable and natural, e.g., "Send an SMS to Amanda telling her the Superman movie is awesome" versus "Hey John, the new Superman movie is great. You should definitely watch it."

Table 2: Corpus statistics.

	Pauses	Turns	Spkrs	Pauses/ Turn	Median Pause Durn
Template	8061	5297	34	1.5	380 ms
Freeform	4409	1590	23	2.7	450 ms

Features

Features were computed in the same manner for both corpora. We studied 15 acoustic features that fall into five groups: (1) pitch trends, (2) duration and intensity, (3) spectral constancy, (4) speaking rate, and (5) periodicity. Although lexical features should be investigated in future work, we limited this study to features that do not require word information, for two reasons. First, such features can be implemented in a client not running full recognition. Second, word-based modeling could be biased due to the specific prompts used, particularly in the case of the Template corpus. All features were computed using only the given utterance and information before the given pause. That is, by design, no look-ahead was permitted within the turn and no other turns from the session or corpus were used for normalization. General descriptions of the feature groups are provided below. Due to space limitations, the implementation of only one sample feature is provided for each group; see also Arsikere et al. (2014).

Pitch trends. F0 (fundamental frequency or pitch) remains fairly steady before nonfinal pauses, but typically falls and/or fluctuates more at utterance ends. This feature type models intonation patterns with the required normalization to account for the spread in F0 'floors' across speakers. A sample feature is ***F0 Drop***: (1) Using the Snack Sound Toolkit (Sjölander, 1997), obtain F0 and the corresponding voiced/unvoiced decision at 10 ms intervals. (2) Divide the utterance into segments that are continuously voiced. (3) Compute the log ratio of the minimum F0 in the last segment to the median F0 over all previous segments.

Duration and intensity. Syllable intensity tends to drop more at utterance ends (because of reduced vocal effort), and continuously-voiced segments tend be longer before nonfinal pauses. This feature type, with a normalization to account for the speakers' 'baseline' intensities, models the above two phenomena. A sample duration feature is ***Continuous Voicing Duration***: (1) Obtain voiced segments using Snack. (2) Compute log ((M-1)*0.01), where M is the number of frames in the last voiced segment (with an inter-frame spacing of 10 ms). A sample intensity feature is ***Intensity Drop***: (1) Compute an intensity (energy) contour using 20 ms frames at 10 ms intervals. (2) Smooth the intensity contour using a 5-point moving-average filter. (3) Detect the peaks in the smoothed contour, and discard those that are within 100 ms of a higher peak; the remaining ones correspond roughly to syllable locations. (4) Compute the log ratio of the final syllable peak to the median of all previous syllable peaks.

Spectral constancy. Speakers tend to maintain a fixed vocal-tract configuration before nonfinal pauses, especially in syllable-final phonemes. This phenomenon can be modelled by analyzing the signal over short time intervals. A sample feature is ***Filter-bank Entropy***: (1) Consider the last 500 ms of the signal. Divide it into 200 ms chunks with 100 ms overlap. (2) *For each chunk*: divide into 20

ms frames with 10 ms overlap; compute the magnitude spectrum of each frame and pass it through a 26-channel Mel filter-bank; find the time variance of each channel; compute the average variance, V_{avg}, across all channels. (3) Find the log minimum of V_{avg} over all chunks.

Speaking rate. Some speakers tend to lower their speaking rate as they approach a nonfinal pause, presumably to gain time to plan content. We model speaking rate using the amplitude modulation of spectral components over long time intervals. A sample feature is ***Intensity Modulation***: (1) Divide the last 1 second of the signal into 300 ms chunks with 100 ms overlap. (2) *For each chunk*: obtain a smoothed intensity contour (as described earlier); compute E_4, the percentage energy above 4 Hz in the amplitude modulation spectrum of the intensity contour. (3) Find the log maximum of E_4 over all chunks.

Periodicity. Voiced segments before nonfinal pauses in our corpora appeared to be more periodic compared to voiced segments at utterance ends. We attribute the aperiodicity in utterance-final syllables to a possible reduction in subglottal pressure, which makes it difficult to sustain regular vocal-fold oscillations. An example of a feature in this class is ***HNR***: (1) Obtain voiced segments using Snack. (2) Divide the last voiced segment into 60 ms frames with 10 ms overlap. (3) Estimate the harmonic-to-noise ratio (HNR) of each frame in the last segment using the algorithm proposed in Qi and Hillman (1997). (4) Compute the 75^{th} percentile of the HNRs over all frames.

Classification experiments

Our task was to determine whether the speaker is done (negative class, since our target class is nonfinal pauses) or not done (positive class), whenever a silence of 100 ms is encountered. We performed leave-one-out cross validation (LOOCV) on each corpus, where one speaker was left out in the training phase (in order to be used for evaluation) and the process was iterated over all the speakers. We also performed cross-corpus evaluation, where one corpus was used for training and the other was used for evaluation. Support vector machines as implemented in LIBSVM (Chang and Lin, 2011) were used as classifiers. All 15 features were scaled to lie in the range [-1,1], prior to training and evaluation. The standard radial-basis-function kernel was used, and the optimal values of C (the penalty parameter) and g (the kernel parameter) were determined via a two-dimensional grid search: C was chosen from $\{2^{-5}, 2^{-3},\ldots,2^{15}\}$ and g was chosen from $\{2^{-15}, 2^{-13},\ldots, 2^{3}\}$.

Results and Discussion

Figure 1 shows the overall ROC curves for LOOCV and cross-corpus evaluation. The baseline cut-off (miss) rates using no additional features are 100% and roughly 36% for response latencies of 100 and 500 ms, respectively. As shown, LOOCV performance using the proposed acoustic features is very similar for the two corpora, achieving a significantly lower cut-off rate of 20.3% in both cases. This suggests that our features are effective for both the slow, deliberate style of the Template corpus and the faster, more spontaneous style of the Freeform corpus.

In the case of the Freeform corpus there is a significant performance difference between LOOCV and cross-corpus evaluation. This shows the importance of training models on spontaneously elicited speech, since personal-assistant interactions in practice are expected be closer to Freeform than Template data. The performance difference between LOOCV and cross-corpus evaluation is smaller in the case of the Template corpus, suggesting that models trained on spontaneously elicited speech would remain effective even if speakers used a slow, deliberate style.

Figure 2 shows ROC curves for a subset of speakers in

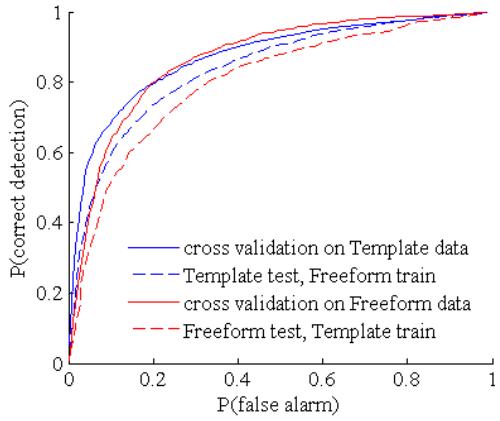

Figure 1: Overall ROC Curves for LOOCV and Cross-Corpus Evaluation.

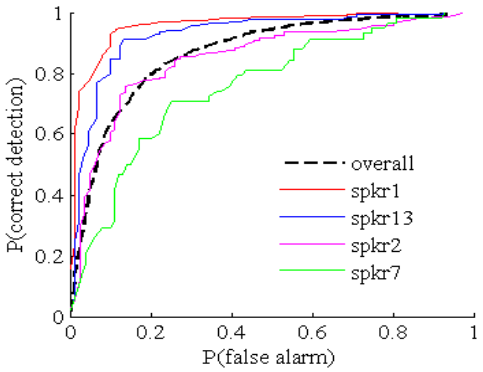

Figure 2: ROC Curves for Sample Speakers in the Freeform Corpus (Results for LOOCV).

the Freeform corpus (for the LOOCV experiment). While speakers differ significantly in absolute performance, they appear to follow the aggregate performance in error trade-offs. Similar trends were observed for the Template corpus. This similarity in overall error trade-offs suggests that the modelling and features used should generalize to new data, even though we expect different speakers and contexts to have significant effects on pausing behaviors.

Feature importance

Despite the similarity in the overall performance across corpora (see Figure 1), we observed differences in the relative contributions of different feature types. It is possible that this is attributable to the different speakers used in the two collections, but we suspect from listening that the differences also reflect the differences in task. As can be seen in Figure 3, periodicity, spectral constancy, and pitch trends showed similar patterns across corpora, with greater absolute importance in the Freeform data.

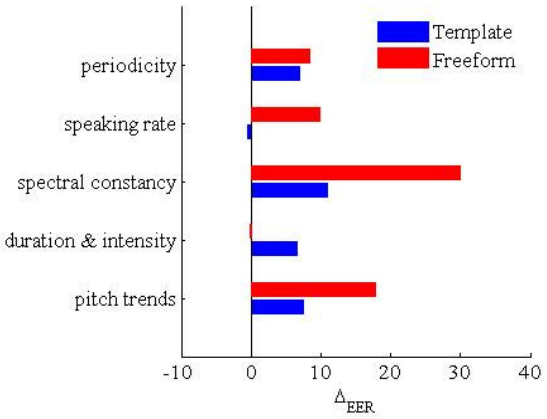

Figure 3: Feature Importance by Type. Importance is Measured as the Change in EER when that Feature Type is Removed from the All-Features Set.

Speaking-rate and duration/intensity features, however, showed a different pattern. Speaking-rate changes helped discrimination in the Freeform data but not in the Template data – perhaps because the latter is already produced at a lower speaking rate. The opposite is true in the case of duration/intensity features, which we plan to investigate individually in the future. Overall, we expect the Freeform results to be a better predictor of real-world performance.

Summary and Future Work

Speech collected using two methods differing in prompt type and task were compared with respect to acoustic features that could predict pause class (nonfinal versus final) from speech preceding the pause. We found almost identical results across corpora for cut-off rate reductions using our prepausal features (from 100% to 20% at a latency of 100 ms), despite much higher rates and lengths of pauses in the less constrained data set. Overall, results show that simple methods for elicitation could be used for modeling, when data with natural pauses is not available in the domain due to the short endpoint thresholds used in conventional systems. Future research includes evaluating lexical features from word recognition, and testing the approach in an interactive system to study performance and user adaptation to enhanced end-of-turn detection.

References

Harish Arsikere, Elizabeth Shriberg, and Umut Ozertem, "Computationally-efficient endpointing features for natural spoken interaction with personal-assistant systems," in *Proc. of ICASSP*, 2014, pp. 3241–3245.

K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma, "Formant-based technique for automatic filled-pause detection in spontaneous spoken English," in *Proceedings of ICASSP*, 2009, pp. 4857–4860.

C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

R. Hariharan, J. Häkkinen, and K. Laurila, "Robust end-of-utterance detection for real-time speech recognition applications," in *Proc. of ICASSP*, 2001, pp. 249–252.

L. Ferrer, E. Shriberg, and A. Stolcke, "Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog," in *Proceedings of ICSLP*, 2002, pp. 2061–2064.

L. Ferrer, E. Shriberg, and A. Stolcke, "A prosody-based approach to end-of-utterance detection that does not require speech recognition," in *Proceedings of ICASSP*, 2003, pp. 605–608.

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, pp. 1526–1540, 2006.

D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech," in *Proceedings of ICASSP*, 1992, pp. 521–524.

Y. Qi and R. E. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *The Journal of the Acoustical Society of America*, vol. 102, pp. 537–543, 1997.

A. Raux and M. Eskenazi, "Optimizing the turn-taking behaviour of task-oriented spoken dialog systems." *ACM Transactions on Spoken Language Processing*, Volume 9, Issue 1, 2012.

K. Sjölander, "The Snack sound toolkit," KTH, (Online: http://www.speech.kth.se/snack/), 1997.