Segmentation of Spoken Dialogue by Interjections, Disfluent Utterances and Pauses

Kazuyuki TAKAGI † *Shuichi ITAHASHI* ‡
e-mail: takagi@cs.uec.ac.jp
† The University of Electro-Communications ‡ University of Tsukuba
† 1-5-1 Chofugaoka, Chofu, Tokyo 182, Japan
‡ 1-1-1 Tennodai, Tsukuba, Ibaraki 305, Japan

ABSTRACT

This paper attempts to segment spontaneous speech of human-to-human spoken dialogues into a relatively large unit of speech, that is, a sub-phrasal unit segmented by interjections, disfluent utterances and pauses. A spontaneous speech model incorporating prosody was developed, in which three kinds of speech segment models and the transition probabilities among them were specified. The segmentation experiments showed that 87.6 % of the segment boundaries were located correctly within 50 msec, 81.2 % within 30 msec, which showed 10.1 point increase in performance comparing with the initial model without prosodic information.

1 Introduction

1.1 Processing unit of spontaneous speech

Segmentation of spontaneous speech is one of the important issues in the creation and the utilization of spoken dialogue corpora. Conversational speech consists of interjectory utterances, hesitations, false starts, unexpected pauses, which break up a sentence into fragments of utterance [1]. In some cases, the last part a sentence and the first part of the following sentence form a single continuous speech interval.

As acoustic features show more variability than laboratory speech, segmentation and labeling of sub-word unit in spontaneous speech are often impractical. A segmentation into a larger speech unit, e.g., a phrasal or a sub-phrasal speech interval bounded by interjections, disfluent utterances and pauses can be considered feasible.

Perhaps one of the simplest and most straightforward ways to classifying segments in spontaneous speech is via a three-state representation in which the segments are (1) pause or silence (P); (2) interjectory utterances (I); and (3) sentence speech (S), which, in this paper, is referred to as all the non-pause, non-interjectory speech segments. Segmentation of conversational speech by these units would be reasonable across various recognition and analysis approaches, though labeling methodologies for smaller unit should differ according to the research purpose. In fact several previous works on spontaneous speech recognition utilize pauses in spontaneous speech or treat inter-pausal phrases and fillers as a unit of processing [2, 3, 4, 5, 6].

1.2 Prosodic characteristics of the units

It has been reported in the literature that interpausal phrases often function as syntactic and intonational units, and that interjectory utterances, filled pauses and hesitations bear distinctive acoustic and prosodic features compared with the surrounding speech segments [7, 8, 9]. An experiment of clustering dialogue speech segments by pitch parameters also implies that prosody would help to distinguish these utterances [10].

2 Speech material

2.1 Spoken dialogue corpora

Two spoken dialogue corpora were used for training the segmentation models and evaluating them. For the creation and training of the initial HMMs, tokens were taken from 10 spoken dialogues in ASJ Simulated Dialogue Corpus [11], which consists of various guidance tasks such as geographic guides, sightseeing guides. The dialogues are not completely spontaneous because they are simulated ones, but are much closer to natural conversation than to read speech. The dialogues were conducted in eye-contacted situation, recorded in three research institutes, spoken by 7 male and 1 female speakers.

Spoken dialogues recorded in the spoken dialogue corpus created by a priority area research project [12] were used as the retraining and the evaluation data set for the segmentation models. Ten dialogues on crossword

solving task, geographical guide, sightseeing guide and telephone shopping were selected, recorded in three research institutions (two of which are different from the training dialogues of ASJ corpus), spoken by 8 male and 4 female speakers. Half of the dialogues are used for retraining and the rest for performance evaluation.

Speech data in both corpora were digitized in 16 kHz, 16 bit. Each dialogue was split into multiple speech files, so that the original dialogue file is rebuilt by simply concatenating them in order. The dialogues last about three to ten minutes.

2.2 Labeling

Segmentation units in this paper are, as described in section 1, P (pause: silence longer than 100 msec), I (interjection), and S(sentence: all the rest of speech). All dialogue speech data were segmented into these three units manually to provide training data and reference labels; the time scale was quantized at 10 msec. Number of tokens in the databases is listed in Table 1. The token data overlapped with an abrupt noise were excluded from the training set. In ASJ Corpus, some segments were overlapped because the dialogue participants' voices are not separately recorded. They were excluded from the training data set. As for the pause data, about a quarter of the tokens were used for training after too noisy token data were removed.

2.3 Feature parameters

The specific phenomena in spontaneous speech studied in this paper are interjections. Interjections are the words that fill pauses, such as "uh", "well", "mmm". They are often uttered at the beginning of a turn or at the major syntactic boundaries of a spoken sentence. It is reported in the literature that interjectory utterances or filled-pauses have distinct phonetic and prosodic characteristics [7, 8, 9]. In this paper, pitch contour information was used for modeling the spontaneous speech as well as phonetic features. Pitch contour was extracted by "Dynamic Pitch Tracker" [13]. Then the first and the second regression coefficients of the contour were calculated after the following error correction and smoothing procedure.

1. deletion of voiced intervals shorter than 50 msec
2. linear interpolation of contour in unvoiced intervals
3. smoothing by Hamming window of 50 msec width

The pitch information was bundled with cepstrum coefficients to provide a feature vector sequence as shown in Table 3.

Table 1: Training and Test Tokens

	Pause (P)	Interjection (I)	Sentence (S)
Initial	291	522	1566
	(2002)	(913)	(1792)
Retraining	215	297	1297
	(860)	(361)	(1314)
Test	1075	152	840

Parenthesized are the total number of tokens in the databases, some of which were removed as inappropriate data for training (as for interjections and sentences).

3 Spontaneous speech model for segmentation

The task is to segment dialogue speech into three kinds of unit, i.e., P (pause), I (Interjection), and S (sentence). As mentioned in section 2.2, each dialogue in the databases is split into multiple files, each of which starts with a pause interval. In order to model spontaneous speech as a sequence of P, I, and S, a probalilistic automaton model in figure 1 was developed.

Each HMM used in the segmentation model is a standard Gaussian mixture density type of left-to-right model, with 3 states and 2 loops, 4 mixture components. The interjection model and the sentence speech model consist of 8 and 10 HMMs respectively, each of whose occurrence probability is assumed to be equal.

Interjectory utterances in the databases were categorized into 7 major interjections, N*, a*, ano*, choQto*, ee*, eeto*, ma*, and "others". The first 7 interjections occupy 83.5 % of all interjection occurrences. Variations of interjections by lengthening of vowels, insertion of a geminate consonant /Q/ and a syllabic nasal /N/, and combinations of these were grouped. "Others" is a model for all the other interjectory utterances. Token data of sentence speech (S) were also clustered into 10 groups simply by the distribution of phoneme occurrences. Pause model consists of a single HMM

Table 2: Recognition rate of each unit

Pause (P)	Interjection (I)	Sentence (S)
99.7	93.5	98.9

Recongition rate [%] of the each unit model, P, I, S as a isolated word recognition.

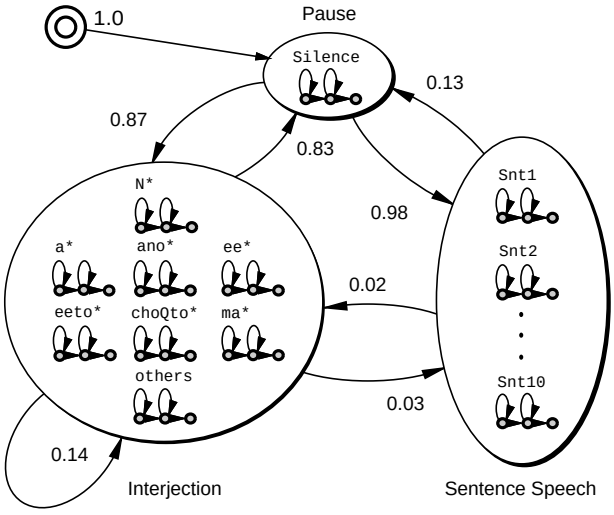

Figure 1: Spontaneous speech model for segmenta¡tion
Interjection model and sentence model consist of several HMMs, whose unigrams are assumed to be equal.

trained with a quarter of pause token data in the training database.

Every HMMs in the segmentation model was first trained by Forward-Backward algorithm using ASJ Corpus data to produce the initial model. Then in the retraining phase, mean vectors of the output probability were updated using half tokens of the test database. Table 2 shows the recognition rate of the three models.

Spontaneous speech was represented by an automaton model with state transition probabilities as illustrated in figure 1. The transition probabilities were calculated from the training database. The model was so constructed that any speech file should start with pause (P) followed by a sequence of P, I, and S.

4 Experiments

4.1 Experimental conditions

Segmentation task was conducted by Viterbi search algorithm using this model as the grammar constraint, in which the scores of the interjection model and the sentence model were given by the best score among the HMMs contained in the model.

Experiments were conduced in four different models, in which each HMM was (1) initial model without pitch parameter, (2) initial model with pitch parameter, (3) retrained model without pitch parameter, and (4) retrained model with pitch parameter. Acoustic analysis conditions of the experiments are shown in table 3.

The mean segmentation error was calculated as the mean value of the distance between the hand-labelled segment boundary and the automatically segmented boundary. Let S_T the total number of correct segments, and S_E the sum of insertions and deletion errors. Segmentation rate was then calculated as $\frac{S_T - S_E}{S_T} * 100$ [%]. The segmentation performances are listed in table 4.

4.2 Results and discussions

The model with pitch information retrained by the test corpus token achieved the best accuracy, by which the insertion error reduced 33 % and the deletion error reduced 30 % compared with the retrained model without pitch information. The segmentation rate of this model was 87.6 % within 50 msec from the hand-labelled boundary, giving 23.7 msec mean error. The model adaptation to the target corpus improved the segmentation rates. The effect appeared more clear when the prosodic parameters were incorporated into the model: the prosodic model improved 10.1 points while the non-prosodic model improved 6.1 points within a tolerance of 50 msec error.

The major improvement was due to the reduction of both deletion and insertion errors of sentence-initial and sentence-internal interjections. Without prosodic information the structure of the utterances with embedded interjection words is ambiguous. The result demonstrates an effect of prosodic information for distinguishing interjections embedded in utterances.

Table 4: Segmentation accuracy

Model	Error [%]		Correct [%]		Mean Error
	Insertion	Deletion	±50 ms	±30 ms	[ms]
Initial[a]	13.8	10.7	75.5	71.1	26.9
Initial$(+F_0)$[b]	12.4	10.1	77.5	68.9	27.3
Retrained[c]	11.3	6.9	81.8	80.1	24.9
Retrained$(+F_0)$[d]	7.6	4.8	87.6	81.2	23.7

a,b Segmentation model created by the initial trainig data (a) without, (b) with pitch information
c,d Segmentaion model retrained by the retraining data of the test database (c) without, (d) with pitch information

Table 3: Experimental conditions

Sampling	16 kHz, 16 bit
Pre-emphasis	$1 - 0.97z^{-1}$
Analysis window	Hamming, 20 msec, 10 msec Δ
Feature parameter	16th order mel cep.
	+ 16th order Δ mel cep.
	+ Δ log power
	(+ F_0's 1st, 2nd regress. coeff.)[b,d]

Superscripts b, d correspond to those in table 4

5 Conclusion

In order to segment spontaneous speech into relatively large units, a spontaneous speech model consisting of pause, interjection and sentence models was proposed, with the transition probabilities among them. Segmentation accuracy was improved significantly by introducing prosodic information.

Future work will elaborate the model by accounting for the occurrence frequency of each interjection model, and will refine the sentence speech models with larger amount of corpus data.

Acknowledgment

This paper used "Dynamic Pitch Tracker"[13] that was developed at Cambridge University.

References

[1] S. Nakagawa and S. Kobayashi, "Phenomena and acoustic variation on interjections, pauses and repairs in spontaneous speech," J. Acoustical Society of Japan, Vol. 51, No. 3, pp. 202 - 210, in Japanese (1995)

[2] J. Hosaka, M. Seligman, and H. Singer, "Pause as a phrase demarcator for speech and language processing," Proc. COLIG 94, pp. 987 - 991 (1994)

[3] T. Takezawa and T. Morimoto, "An efficient predictive LR parser using pause information for continuously spoken sentence recognition," Proc. ICSLP 94, pp. 1 - 4 (1994)

[4] J. Murakami and S. Matsunaga, "A spontaneous speech recognition algorithm using word trigram models and filled-pause procedure," Proc. ICSLP 94, pp. 819 - 822 (1994)

[5] F. Ehsani, K. Hatazaki, J. Noguchi, and T. Watanabe, "Interactive speech dialogue system using simultaneous understanding," Proc. ICSLP 94, pp. 879 - 882 (1994)

[6] K. Itou, T. Akiba, S. Kamijo and K. Tanaka, "A dialogue processing method based on interpausal utterance," Tech. Rep. Information Processing Society of Japan, Vol. 95, No. 73, pp. 135 - 138, in Japanese (1995)

[7] D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech," Proc. ICCASP 92, pp. I-521 - 524 (1992)

[8] E. E. Shriberg and R. J. Lickley, "Intonation of clause-internal filled pauses," Proc. ICSLP 92, pp. 991 - 994 (1994)

[9] M. Kawamori, T. Kawabata and A. Shimazu, "A phonological study on Japanese discourse markers (II)," Tech. Rep. Information Processing Society of Japan, Vol. 95, No. 120, pp. 13 - 20 (1995)

[10] K. Takagi and S. Itahashi, "Prosodic pattern of utterance units in Japanese spoken dialogues," Proc. ICSLP 94, pp. 143 -146 (1994)

[11] "Continuous Speech Corpus for Research," Vol. 7, CD-ROM, Acoustical Society of Japan (1993)

[12] "Simulated Spoken Dialogue Corpus," Vol. 7, CD-ROM, by "Research on Understanding and Generating Dialogue by Integrated Processing of Speech, Language and Concept," Grant-in-Aid for Scientific Research in Priority Areas by MESC Japan (1995)

[13] T. Robinson, "Dynamic pitch tracker," Cambridge University Engineering Department, ftp://svr-ftp.eng.cam.ac.uk /pub /comp.speech /analysis/ (1992)