CONVERSATION DETECTION IN AMBIENT TELEPHONY

Aki Härmä and Kien Pham

Philips Research Laboratories, Eindhoven, The Netherlands

ABSTRACT

In some speech communication applications such as distributed hands-free telephony it is important that the system can detect the conversational state of a call. This cannot be performed by speech activity only because the captured signal may also contain conversation between two local people, or additional speech noise sources such as speech sounds from a radio or television. In this paper we compare known algorithms and introduce a new algorithm for the real-time detection of active conversation between an incoming caller and a local user. The method is based on the mutual information in speech activity, detection of back-channel speech activity, and statistics of overlapping speech. The proposed method gives over 90% accuracy within one minute observation period which is a clear improvement over the performance of earlier techniques.

Index Terms— Conversation detection, ambient telephony, speakerphone

1. INTRODUCTION

In all speech communication systems it is obvious when there is a line open between two telephone devices. However, having a call open does not necessarily imply that there is a conversation going on between two or more individuals located at the different ends of the communication line. In this paper we address the problem of dynamic detection of conversational activity between two people. In this paper the focus is in the use of conversation detection in *ambient telephony* where calls may have fluctuating activity level and there may be multiple local and remote participants [1]. For example, the ambient telephone system shown in Fig. 1 consists of five terminal devices, or *phonelets*, distributed in the home environment. One of the phonelets is the master phone which is the gateway to different transmission platforms such as IP, or a cellular phone network. The network of phonelets is used as a combined loudspeaker/microphone array to perform the sound reproduction and capture in a spatially selective way such that the system creates a virtual speaker-phone representing each caller in some desired location in the users environment. The speech reproduction in the network of phonelets can be driven in such that the incoming caller appears moving smoothly with a tracked local user, for example, from one room to another, or there are multiple incoming caller positioned in different places.

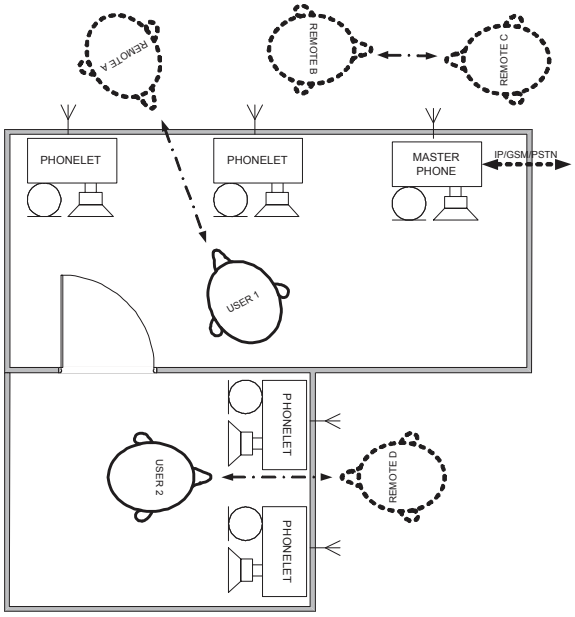

Fig. 1. Ambient telephone system with multiple users and calls.

In the scenario shown in Fig. 1 User 1 is in active conversation with a remote person (Remote A). However, there is also another call open simultaneously which has been rendered to the right hand side of the room. There is no active conversation between User 1 and the remote people on the right hand side (B and C). However, at the same time there is also a conversation going on between User 2 and Remote D in the other room. For the ambient telephone system, there are three equivalent open calls in the system. If User 1, for example, wants to mute the voices of Remote people B&C, or to go to the second room to have an *ex tempore* conference with User 1, User 2, and the remote caller A, it is not immediately clear to the system which callers should be muted or moved. For a user, it is natural that the user interaction such as muting a call, moving a call to another location, or closing a call should address the current active conversation of the user.

The most of obvious indicator for conversation is that the talkers take turns such that when one person is talking, the

other one is listening. Brady [2] demonstrated that turn-taking in telephone conversation can be modeled efficiently with a six-stage state transition model for the conversation. The detection of conversation based on mutual information (MI) computed from speech activity sequences was first proposed by Basu [3]. The MI was computed between two binary audio signals, which contained the voice activity decisions of the speakers and classified them into a pairwise conversation if the value exceeded the predefined threshold. A similar approach was also introduced in [4], where they constructed a single Hidden Markov Model (HMM) with different group settings for modeling the turn-taking dynamics. While earlier papers focused on detecting conversation from an entire recording, the current application requires a dynamic detector which is capable to detect the conversational state preferable within one minute from the start of the conversation. In practice, it would be desirable to detect a conversation with at least 90% precision within one minute from the start of the conversation.

The largest problem in conversation detection is that people talk on simultaneously. This behavior can be seen more frequently in a telephone conversation compared to a face-to-face interaction as the turn-taking relies mostly on the verbal cues. For example, in the Spoken Dutch Corpus for telephone dialogues approximately 41% of the turn-taking exchanges contain overlapping speech [5]. The results of this paper show that the presence of overlapping speech significantly reduces the performance of a pure MI-based conversation detection algorithm. One way to improve the performance is to incorporate models of overlapping speech in the conversation model. In this paper we focus on two aspects of overlapping conversational speech: back-channel (BC) activity and statistics of overlapping speech.

Back-channel speech activities ('yeah!', 'right!, 'mmmm..') is a natural part of conversation in most languages. For example, in the Dutch dialogue database mentioned above 19% of temporally overlapping talk can be interpreted as back-channel activity. Secondly, the amount of overlap in speech is an important cue for conversation detection.

2. CONVERSATION DETECTION MODEL

In the conversation model M *local* and N *remote* speech signals are represented by sequences of S-dimensional features computed from the incoming signals. It is necessary to follow the activity over W frames and therefore it is convenient to stack the feature vectors into matrices corresponding to an observation period in the following way

$$\mathbf{V}_{lm} = [\mathbf{v}_{lm}(k-W+1)\cdots\mathbf{v}_{lm}(k)] \quad (1)$$

$$\mathbf{V}_{rn} = [\mathbf{v}_{rn}(k-W+1)\cdots\mathbf{v}_{rn}(k)]] \quad (2)$$

where the feature vectors at frame number k are denoted $\mathbf{v}_{rn}(k)$ and $\mathbf{v}_{lm}(k)$, with $n = 0,\cdots,N-1$ and $m = 0,\cdots,M-1$, respectively. For user m and N incoming calls the problem of conversation detection is a task of finding the $\hat{n}(k)$th remote caller such that

$$\hat{n}_m = \text{argmax}_n[C(\mathbf{V}_{lm}, \mathbf{V}_{rn})] \quad (3)$$

where $C(\mathbf{V}_{lm}, \mathbf{V}_{rn})$ is the likelihood of conversation between talkers m and n. Indices n and m are dropped from the subsequent equations because the focus is on pair-wise analysis and we use symbols l and r, for a local and a remote talker, respectively. Generally, a feature vector $\mathbf{v}$ could contain for example Mel-Frequency Cepstral Coefficients (MFCCs), or pitch estimates. An estimator C based on raw speech features could be trained for detect conversation but this may be difficult because of the large variance in the feature values and the fact that a large part of the variability is irrelevant for the task of detecting a conversation. One of the most important characteristics of conversational speech is turn-taking, which in an ideal case can be seen in synchronous alternation in the speech activity between the two talkers. Also, it is known that there are systematic pitch changes related to the conversational structure. Therefore, the feature vector $\mathbf{v} = [s, f]$ computed in the proposed model consists of the speech activity s and the fundamental frequency, or pitch, of voiced speech f. The set of features is computed in each time frame of 512 samples at the audio sampling rate of 22.05 kHz. The speech activity detector is that of the standardized AMR-WB coder and pitch estimation was based on the well-known YIN algorithm.

2.1. Detection based on speech activity

In ideal conversational turn-taking the binary speech activity sequences of the two talkers, r and l, are complementary, that is, $s_l(k) = 1 - s_r(k)$, where, $k = 0,\cdots,W-1$, or $\mathbf{s}_l = 1 - \mathbf{s}_r$ in the vector notation. The normalized cross-correlation (NCC) function $R_{s_l s_r}(p)$ between these sequences would give -1 for the zero lag. Therefore a *conversation likelihood* estimator in $[0, 1]$ based on finding the minimum of the NCC can be defined as follows

$$C_{\text{ncc}}(\mathbf{s}_l, \mathbf{s}_r) = (1 - \min[R_{s_l s_r}(p)])/2, \quad (4)$$

where p is limited to some region R around the zero-lag correlation. In initial experimentation it was found that it is beneficial to low-pass filter the speech activity sequences before the computation of the NCC function such that only fluctuations of speech activity below approximately four Hertz are taken into account. The conversation detection results computed over 19 conversation sequences from the IFA database are shown in the top panel of Fig. 2. In the example, a pair of speech activity sequences was labeled conversation if $C_{\text{ncc}}(\mathbf{s}_l, \mathbf{s}_r) > 0.4$. The false conversations were formed by comparing two talkers from different conversations of the database. The precision, accuracy, and recall values (see caption of Fig. 2) improve when the duration of the observation

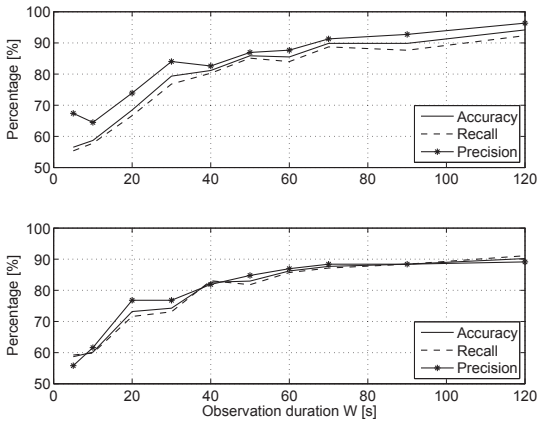

Fig. 2. The performance of NCC-based (top) and MI-based (bottom) conversation detectors as a function of the length of the observation. Accuracy represents the proportion of correct classifications in the population. Precision is the probability of classifying the relevant item in the population. Recall is the probability that the classified item is relevant in the sample.

window is increased. The mutual information (MI) has been used earlier in the place of the cross-correlation metrics [3]. The detector based on MI can be defined in the following way

$$C_{\mathrm{mi}}(\mathbf{s}_l, \mathbf{s}_r) = \sum P(s_l, s_r) log \frac{P(s_l, s_r)}{P(s_l)P(s_r)} \qquad (5)$$

where the $P(s_l, s_r)$ are the probability distributions which are in practice approximated by two-dimensional histograms of $\mathbf{s}$ values computed over the observation window of W feature vectors. The $P(\mathbf{s}_l)$ and $P(\mathbf{s}_r)$ are correspondingly the marginal distributions computed from $P(\mathbf{s}_l, \mathbf{s}_r)$. The detection results of C_{mi} are shown in the bottom panel of 2. In some studies, e.g., [6], the MI-based conversation detection has given higher accuracy that the use of a correlation function. The current does not show a clear difference between C_{mi} and C_{ncc}. However, the results of C_{mi} and C_{ncc} at the observation period of one minute is around 85% in both methods.

2.2. Proposed model

The block diagram of the conversation detection model C proposed in this paper is shown in Fig. 3. The model has three main components: a back-channel detector (BC), MI estimator, and an overlap statistical analyzer (OLS). The back-channel BC activity decreases the C_{mi} likelihood in the case of a real conversation. Therefore, a specific algorithm was developed to detect and remove the back-channel activity from $\mathbf{s}$ prior to the computation of the C_{mi}.

The back-channel speech activity consists of short isolated utterances which are simultaneous with the speech of the

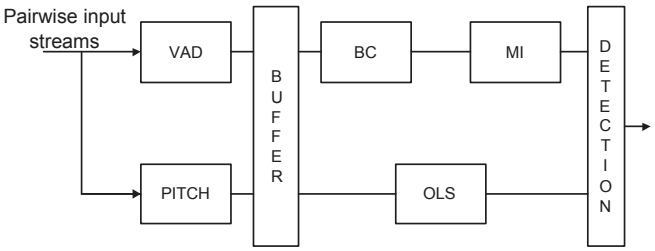

Fig. 3. The block diagram of the proposed method

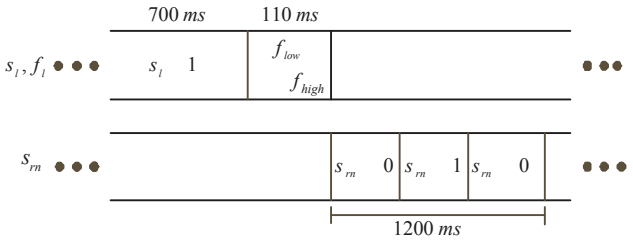

Fig. 4. Back-channel detection method.

other person. Such segments can be detected relatively easily from speech activity sequences $\mathbf{s}$. However, it was found that the plain removal of all short isolated utterances actually led to a significant increase of the mutual information in the false cases, and therefore, a worse overall performance. In order to to eliminate only true back-channel activity a model using pitch information was developed. It is known that there are systematic pitch effects in the speech of the *active* talker before the back-channel activity of the listener. The most prominent effects are the lowering of the pitch to *invite* back-channel affirmative from the listener [7], and the rising pitch in questions. The back-channel detection model proposed in this paper is a rule-based template-matching method illustrated in Fig. 4 based on the description given in [7]. The analysis of a window of W feature vectors starts with a search for a pitch *trigger* from a continuous speech segment of Talker A. If the pitch is below a lower f_{low} or above an upper frequency limit f_{high} in a segment after at least 700 ms of continuous speech, and Talker B is silent during the segment, the part is selected as a trigger. The frequency limits are determined as 28th and 75th percentile from the histogram of pitch values of that talker. When a potential *trigger* is found, the method searches for an isolated speech utterance within a 1.2s window from the speech activity sequence of Talker B after the time of the trigger. This is performed for each pair of talkers in both ways.

The speech activity values $s(k)$ in segments detected as back-channel activity are then set to zero. This gives the back-channel free voicing feature sequences denoted by $\hat{\mathbf{s}}$. The back-channel free MI-based conversation detector is given by $C_{\mathrm{bfmi}}(\mathbf{s}_l, \mathbf{s}_{rn}) = C_{\mathrm{mi}}(\hat{\mathbf{s}}_l, \hat{\mathbf{s}}_{rn})$.

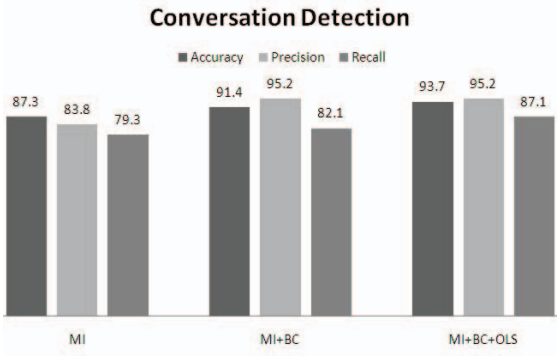

Fig. 5. The results from the conversation detections in one minute observation period.

It is known that in a natural dialogues both the speaker and the listener contribute to minimize the speech overlaps and aim for a smooth transitions in turn-takings. Therefore using some temporal statistics it is possible to support the decision and improve the reliability in uncertain cases. One alternative is to use the proportion of the overlap speech activities between the cross pair signals. The overlap ratio is given by

$$C_{\text{overlap}} = \frac{P(\mathbf{s}_l = 1, \mathbf{s}_r = 1)}{\max(\sum \mathrm{P}(\mathbf{s}_\mathrm{l} = 1), \sum \mathrm{P}(\mathbf{s}_\mathrm{r} = 1))} \qquad (6)$$

where $P(\mathbf{s}_l = 1, \mathbf{s}_r = 1)$ is the joint distribution of the overlapped speech and both $P(\mathbf{s}_l = 1)$ and $P(\mathbf{s}_r = 1)$ correspond to the marginal distributions of speech activity of each talker. The analysis of conversational data suggests that if over 39% of the speech activities are overlapping, then it is unlikely that they are in the same conversation. The final detection, see Fig. 3, is then based on the following rule: $C_{\text{detection}} = \min(\mathrm{C}_{\text{bcmi}}, \mathrm{C}_{\text{overlap}})$.

3. EXPERIMENTS

The method was evaluated with seven 15 minute conversations from the Dutch IFA corpus [8]. We simulated our method dynamically with 60s window over the data and evaluated the efficiency of each component with 3 different setups. The results in terms of accuracy, precision and recall are shown in Fig. 5. In general, the use of back-channel detection and temporal statistical analysis shows clear improvements in accuracy, precision and recall. Most noticeable improvement is achieved in the precision, which results from the dialogue scenes where one speaker is strongly holding the floor and the other only producing back-channel feedbacks or overlapped responses. In these cases using only the MI scores cause the real conversations to remain undetected. The precision values for the proposed model are above the 90% target set for the one minute observation time.

4. CONCLUSIONS

Distributed hands-free telephone systems can benefit in many ways from computational models that can separate active conversation of two users from other combinations of simultaneous speech signals. In this paper we have introduced the problem and a novel algorithm which can detect a real conversation from other types of speech signal combinations in a real-time communication setting. The method is based on a back-channel activity model, mutual information in speech activity sequences, and statistics of overlapping speech. The proposed algorithm gives over 90% accuracy in one minute observation period with a database of Dutch conversations. The results are significantly better than using the plain correlation or mutual information metrics earlier proposed in the literature, and also tested in the current paper. However, it is possible that a part of this improvement is specific for the Dutch language conversation data used in the experiments. Therefore, the experiment should be repeated in the future with a broader selection of conversation data.

5. REFERENCES

[1] A. Härmä, "Ambient telephony: scenarios and research challenges," in *Proc. INTERSPEECH 2007*, Antwerp, Belgium, August 2007.

[2] P. T. Brady, "A model for generating on-off speech patterns in two-way conversation," *Bell Syst. Tech. J.*, pp. 2445–2472, September 1969.

[3] S. Basu, *Conversational Scene Analysis*, Ph.D. thesis, MIT,Cambridge, 2002.

[4] O. Brdiczka, J. Maisonnasse, and P. Reignier, "Automatic detection of interaction groups," in *ICMI*, Trento, Italy, 2005.

[5] L. Bosch, N. Oostdijk, and J. P. De Ruiter, "Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues," in *International Conference on Text Speech and Dialogue*, 2004.

[6] D. Wyatt, T. Choudhury, and J. Bilmes, "Conversation detection and speaker segmentation in privacy-sensitive situated speech data," in *INTERSPEECH*, Antwerp, Belgium, 2007.

[7] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[8] R. van Son, W. Wesseling, E. Sanders, and H. van den Heuvel, "The IFADV corpus: A free dialog video corpus," in *International Conference on Language Resources and Evaluation*, Marrakech, Marocco, 2008.