NEW FEATURES FOR IMPROVING VAD WHEN DEALING WITH FAR-FIELD AND MULTI-SPEAKER SPEECH

Oscar Varela Serrano [1], Rubén San-Segundo Hernández [2], Luis Alfonso Hernández Gómez[3]

[1] Telefónica I+D. Madrid, Spain
[2] Grupo de Tecnología del Habla, UPM
[3] Grupo de Aplicaciones del Procesado de Señal, UPM
ovs@tid.es

ABSTRACT

This paper describes new acoustic features for improving VAD (Voice Activity Detection) when dealing with speech mixed with far-field and multi-speaker speech. Background voices are one of the major causes for the degradation of speech recognition performance in spoken dialog systems (specially over mobile phones). Also, in any audio indexing application, to separate the voice of a target speaker from other background speakers can be necessary. This paper studies three new features to discriminate between near-field, far-field and background multi-speakers speech: 1) the percentage of frame-by-frame change for the best HMM mixture in a HMMs-based VAD; 2) the Mahalanobis distance between MFCCs from consecutive speech frames, and 3) the maximum auto-correlation value for each speech frame. Experimental results on the Av16.3 speech database for the best feature, obtain classification errors below 19% for near-field vs. far-field speech, and 3.5% for one-speaker vs. multi-speaker.

Index Terms: VAD, far-field speech, multi-speaker speech.

1. INTRODUCCIÓN

This paper addresses the problem found in many speech-based applications when speech of the user to be recognized is contaminated with background voices from other speakers standing still or moving. Far-field speech is specially problematic and usual in mobile phone scenarios, where the main speaker can be situated in open environments surrounded with far-field interfering speech from other speakers. In this case, VAD systems can detect far-field speech as coming from the user increasing the speech recognition error rate. Generally, errors caused by background voices mainly increase word insertions and substitutions, leading to important dialogue misunderstandings.

In several previous works, similar measures as the ones this work considers have been used for dereverberation techniques. In [1] for example, authors use the idea of reverberation for restoring speech degraded by room acoustics using stereo (two microphone) measures. To do this, cepstra operations are made when observations have nonvanishing spectra. Other dereverberation technique, presented in [2], uses the pitch as primary analysis feature. That method starts estimating pitch and harmonic structure of the speech signal to obtain a dereverberation operator. After that, this operator is used to enhance the signal through an inverse filtering operation. Single channel blind dereverberation was proposed in [3] based on auto-correlation functions of frame-wise time sequences for different frequency components. A technique for reducing room reverberation using complex cepstral deconvolution and the behavior of room impulse responses was presented in [4]. Reverberation reduction using least square inverse filtering has been also used to recover clean speech from reverberant speech. Yegnanarayana shows in [5] a method to extract time-delay between two speech signals collected at two microphone locations. The time-delay is estimated using short-time spectral information (magnitude, phase or both) based on the different behavior of the speech spectral features affected by noise and reverberation degradations. Finally, Cournapeau shows in [6] a VAD based on High Order Statistics to discriminate close and far-field talk, enhanced by the auto-correlation of LPC residual.

Nowadays there is an increasing interest on the relevance of VAD systems in real applications. New VAD techniques are being proposed, see for example the work of Ramirez et al. [7] on robust VAD using the Kullback-Leibler divergence measure. However, although experimental results are usually given for the AURORA database, to our knowledge there are no similar results for speech in the presence of far-field voices.

In this paper, trying to contribute to the improvement of VAD systems in the presence of background speech, we present a preliminary analysis of new features suitable to classify near-field, far-field and multi-speaker speech. We consider simple acoustic feature that could be easily and cost-effective integrated in state-of-the art VAD.

The rest of this paper is organized as follows: the speech database and our experimental evaluation framework are described in Section 2. Section 3, 4 and 5 present three different features for far-field and multi-speaker discrimination together with their corresponding evaluation results. Finally, some conclusions are given in Section 6.

2. SPEECH DATABASE

The database we used in this work is the Av16.3 speech database composed of audio-visual data recorded in a meeting room context. For this work, only the audio data has been considered. This audio has been recorded with 16 microphones perfectly synchronized and calibrated conveniently. For each recording, there are 16 audio WAV files from the two circular 8-microphone arrays (Fig. 1) sampled at 16 KHz and WAV files recorded from lapels also sampled to 16 KHz. It is specially important to point out that overlapped speech has been recorded when there are several speakers speaking simultaneously.

In order to allow for such a broad range of research topics, "meeting room context" is defined in a wide way. This includes a high variety of situations, from "meeting situations" where speakers are seated most of the time, to "motion situations" where speakers are moving most of the time (Fig.1). Audio files are named in function of the speakers characteristics (for more details see [8]). These files have been resampled down to 8 Khz (for simulating a telephone channel) and randomly divided into three sets: training (80%), validation (10%) and test (10%). The feature analysis has been performed over the training set.
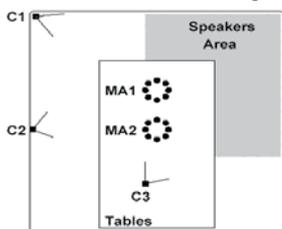
Figure 1. *MA1 and MA2 8-microphone circular array. See Speakers Area. This figure has been obtained from [8].*

3. PERCENTAGE OF CHANGES FOR THE BEST MIXTURE IN A HMM-BASED VAD SYSTEM

This section presents a study about the discrimination power between near-field vs. far-field speech using as feature the percentage of times the best mixture (in a maximum likelihood sense) of a speech HMM model in a HMM-based VAD change across a set of successive frames. Our VAD system uses two one-state HMMs (noise and speech models) including 200 Gaussian mixtures. This high number of components in the Gaussian mixture introduces more mixture variability producing higher frame-to-frame best mixture variability for multispeaker signals. The VAD

system uses a MFCC vector (generated from a 12 Mel filter-bank analysis) formed by the first 8 cepstrum coeficients, normalized energy and delta energy. The HMMs models have been trained by means of Baum Welch re-estimation (ec. 1).

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} N(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s} \quad (1)$$

where M_S is the number of mixture components in stream s, c_{jsm} is the weight of the m´th component and $N(\bullet; \mu; \Sigma)$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ.

Initially the best mixture was selected after applying its mixture weight but small Gaussian variability was found. In order to increase this variability, mixtures weights were removed from the best mixture computation. In this case, a lot of candidates of winner gaussians were obtained when processing all frames. The measure we propose is the percentage of changes of the best Gaussian along N consecutive frames.

Figures 2 and 3, show the distribution of the percentage of changes considering N=100 and N=1000 frames respectively. Only speech frames are considered in this study. The noisy frames are discarded.
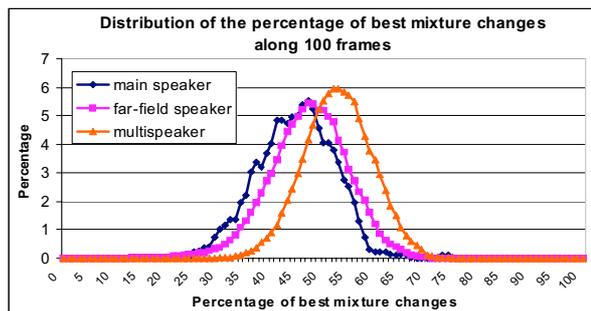
Figure 2. *Distribution of the percentage of changes considering N=100 frames.*
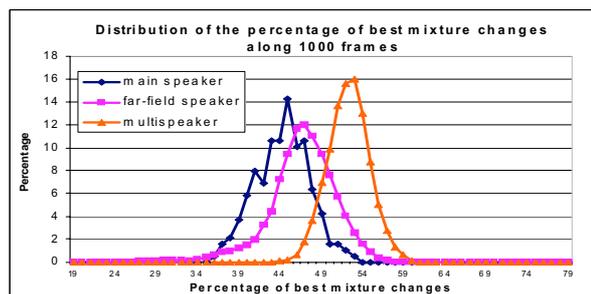
Figure 3. *Distribution of the percentage of changes considering N=1000 frames.*

As it is shown in figures 2 and 3, the percentage of best mixture changes is higher for speech coming from several speakers at the same time. This feature can discriminate very well between main speaker and multispeakers voices. In this case, the error is lower than 26% and 10% for 100 and 1000 frames respectively. When the number of frames considered for the percentage computation (N) increases from 100 to 1000 the measure is better estimated and the

discrimination power increases. On the other hand, the discrimination power between main speaker and far-field speaker voices is not good enough. Anyway, it is possible to see a higher percentage of changes for far-field speech.

4. MAHALANOBIS DISTANCE BETWEEN MFCCs

This feature consists of computing the Mahalanobis distance between MFCC vectors obtained from consecutive speech frames. Every vector contains the first 8 MFCC coeficients, normalized energy and delta energy. Mahalanobis distance, ec. (2), is used to evaluate the similarity between multidimensional random variables:

$$d_M(\vec{x}_{i.};\overline{x}_j) = \sqrt{(\vec{x}_{i.} - \overline{x}_j)S^{-1}(\vec{x}_{i.} - \overline{x}_j)^t} \quad (2)$$

where S is the covariance matrix of the variable vector $(x_1, x_2, ..., x_k)$. The distributions of Mahalanobis distance between consecutive frames for the main speaker, far-field speaker and multi-speaker speech are shown in figure 4. Figure 4 shows the histogram of the Mahalanobis distances between consecutive frames. As it is shown, main speaker speech presents lowest distance while multi-speaker presents the highest ones. At this point, the analysis were extended to groups of N frames, considering N=50 and N=500 frames. Again only speech frames were considered and noisy frames were discarded. In this process, the minimal distance along N consecutive frames is computed. Figures 5 and 6 shows the distributions of the minimal distance.
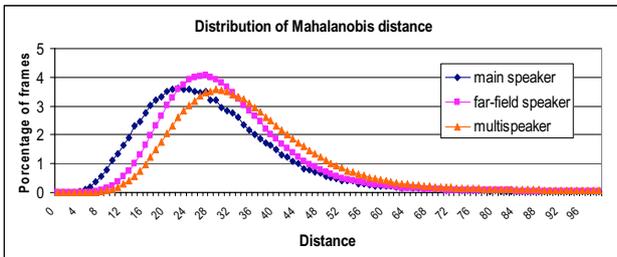

Figure 4. *Distribution of Mahalanobis distance distributions for main speaker, far-field speaker and multi-speaker speech.*

As it is shown in figures 5 and 6, the minimal distance along the N frames is higher for speech coming from several speakers at the same time. This feature can discriminate very well between main speaker and multispeakers voices. In this case, the classification error is lower than 24% and 14% for 50 or 500 frame segments respectively. When the number of frames considered for the minimal computation (N) increases from 50 to 500 the minimum is better estimated and the discrimination power increases.

The discrimination power between main speaker and far-field speaker voices with this feature is better compared to the previous feature. In this case, errors are lower than 35% and 27% for 50 or 500 frames. Other related measures, like the maximum, average, variance

or kurtosis of the Mahalanobis distance, were also tested, but only the minimum distance showed an interesting relationship with the voice type. We think this is due to the fact that a low minimum distance is obtained during stationary speech zones: very infrequent in far-field and multi-speaker speech.
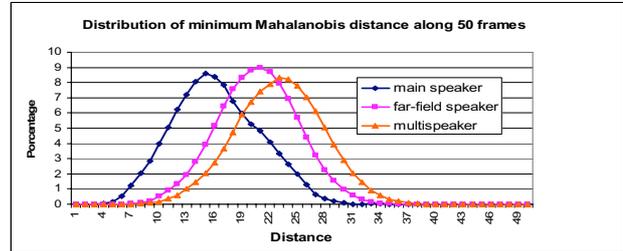

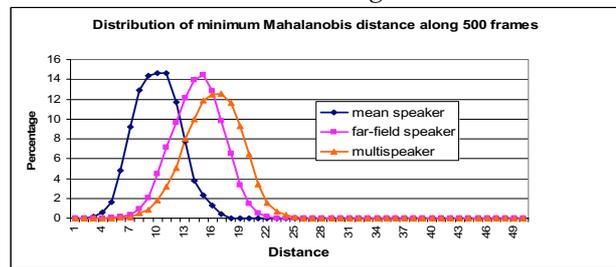**Figure 5**. *Distribution of minimum Mahalanobis distance considering N=50.*

Figure 6. *Distribution of minimum Mahalanobis distance considering N=500.*

5. MAXIMUM AUTO-CORRELATION OBTAINED WHEN COMPUTING THE PITCH

In this case, the study was focused on the behavior of the auto-correlation values when computing the pitch at every frame. Considering only voice frames, the maximum of auto-correlation in pitch regions is considered. Fig. 7 presents the maximum auto-correlation distributions for main speaker, far-field speaker and multispeaker speech.
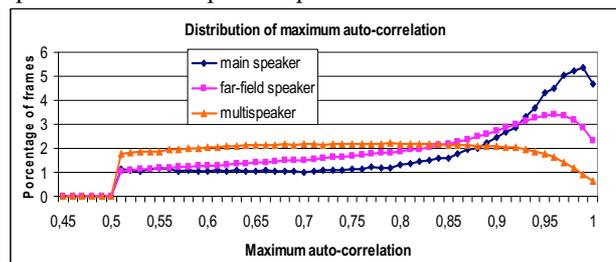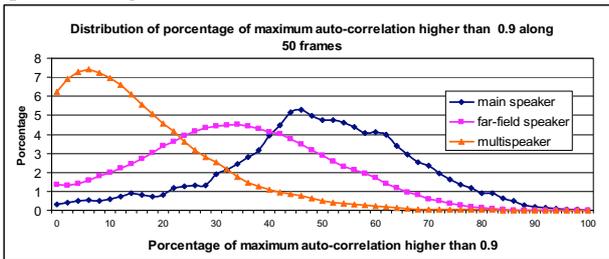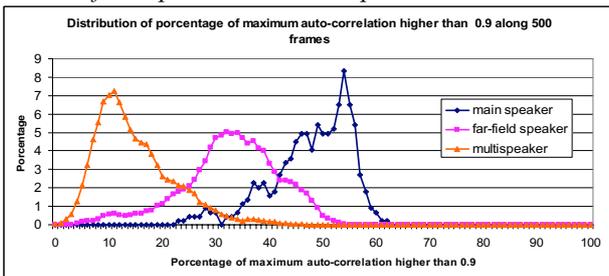

Figure 7. *Distribution of maximum auto-correlation for main speaker, far-field speaker and multispeaker.*

Fig 7. shows very different behaviors for the maximum auto-correlation value in the three cases, specially for auto-correlation values higher than 0.9. There are many more frames in the case of the main speaker speech and very few in the case of multi-speaker speech. So after considering this effect, the percentage of frames (along N frames) with a maximum auto-correlation higher than 0.9 was computed for the three types of speech. Figures 8 and 9 show the distributions of the percentage of maximum auto-correlation values higher than 0.9 for

main speaker, far-field speaker and multi-speaker speech along 50 and 500 frames.

Figure 8. *Distribution of porcentage of maximum auto-correlation higher than 0.9 for main speaker, far-field speaker and multi-speaker N=50.*

Figure 9. *Distribution of porcentage of maximum auto-correlation higher than 0.9 for main speaker, far-field speaker and multi-speaker N=500.*

As it is shown in figures 8 and 9, the percentage along the N frames is lower for speech coming from several speakers at the same time. This feature can discriminate very well between main speaker and multispeakers voices. In this case, the error is lower than 15% and 3.5% for considering 50 or 500 frames respectively. When the number of frames considered (N) increases from 50 to 500 the percentage is better estimated and the discrimination power increases. The discrimination power between main speaker and far-field speaker voices is better compared to the previous two features. In this case, classification errors are lower than 33.5% and 19% for 50 and 500 frames respectively.

6. CONCLUSIONS

This paper presents new successful features for improving VAD (Voice Activity Detection) when main speaker speech is mixed with far-field and multi-speaker speech. Generally, these features can be used to improve the behavior of any application in which it is necessary to discriminate the main speaker speech from far-field speech and multi-speaker speech. This study has been done with the Av16.3 speech database but the audio files have been resampled to 8Khz in order to simulate a telephone channel. The first feature proposed has been the percentage of changes of the mixture with the maximum likelihood, considering a VAD system based on HMMs. Results show better performance for multi-speaker speech rejection.

The second one was the Mahalanobis distance between the MFCCs of consecutive speech frames. In this case, the results were better than previous feature ones. Error between main speaker speech and multi-

speaker speech was lower than 24% and 14% for considering 50 or 500 frames respectively. On the other hand, comparing main speaker speech vs. far-field speech, classification error was lower than 35% and 27% for 50 and 500 frames.

Finally, the best feature has been the maximum auto-correlation value obtained when computing the pitch at every frame. This feature can discriminate very well between main speaker and multi-speaker voices. Although some measures over this maximum has been processed, percentage of frames with a maximum of auto-correlacion value higher than 0.9 is the one which gets the best results. In this case, the error was lower than 15% and 3.5% for considering 50 or 500 frames respectively. When comparing main speaker speech and far-field speech, classification errors are lower than 33.5% and 19% for 50 and 500 frames.

For all the features, when the number of consecutive frames considered for feature computation increases, the discrimination power increases. It is important to remark that a good performance for real time applications is obtained for the second and third features whose behavior when considering 50 frames is very good.

7. REFERENCES

[1] Petropulu, A. P., and Subramaniam, S.,"*Cepstrum based deconvolution for speech dereverberation*", IEEE Trans. Speech and Audio Proc., pp. 9-12, 1994.

[2] Nakatani, T. and Miyoshi, M.,"*Blind dereverberation of single channel speech signal based on harmonic structure*", pp. 92-95, ICASSP 2003.

[3] Ohta, K. and Yanagida, M., "*Single channel blind dereverberation based on auto-correlation functions of frame-wise time sequences of frequency components*", Iwaenc 2006 – Paris – September 12-14, 2006.

[4] Bees, D., Kabal, P., and Blostein, M., "*Application of complex cepstrum to acoustic dereverberation*", Proc. Biennial Symp. Commun. (Kingston, ON), pp. 324-327, June 1990.

[5] Yegnanarayana, B., Mahadeva Prasana, S. R., Duraiswami, R. and Zontkin, D., "*Processing of Reverberant Speech for Time-Delay Estimation*", IEEE Trans. Speech and Audio Proc., pp. 1110-1118, vol. 13, nº 6, 2005.

[6] Cournapeau, D. And Kawahara, T., "*Evaluation of Real-Time Activity Detection based on High Order Statistics*", pp. 2945-2948, Interspeech 2007.

[7] Ramírez, J., Segura, J., Benítez, C. and Rubio, A., "*A New Kullback-Leibler VAD for Speech Recognition in Noise*", IEEE Signal Proc., vol 11, nº 2, pp. 266-269, 2004.

[8] AV16.3: *An Audio-Visual Corpus for Speaker Localization and Tracking*.