Robust voice activity detection in stereo recording with crosstalk

Prasanta Kumar Ghosh, Andreas Tsiartas, Panayiotis Georgiou, and Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory,
Department of Electrical Engineering, University of Southern California,
Los Angeles, CA 90089

`prasantg@usc.edu, tsiartas@usc.edu, georgiou@sipi.usc.edu, shri@sipi.usc.edu`

Abstract

Crosstalk in a stereo recording occurs when the speech from one participant is leaked into the close-talking microphones of the other participants. This crosstalk causes degradation of the voice activity detection (VAD) performance on individual channels, in spite of the strength of the crosstalk signal being lower than that of the participant's speech. To address this problem, we first detect speech using a standard VAD scheme on the merged signal obtained by adding the signals from two channels and then determine the target channel using a channel selection scheme. Although VAD is performed on a short-term frame basis, we found that the channel selection performance improves with long-term signal information. Experiments using stereo recordings of real conversations demonstrate that the VAD accuracy averaged over both channels improves by 22% (absolute) indicating the robustness of the proposed approach to crosstalk compared to the single channel VAD scheme.

Index Terms: voice activity detection, two channel recording, crosstalk

1. Introduction

Detection of speech and non-speech portions from conversations involving multiple speakers such as multi-channel audio recorded at meetings has become a critical initial step for many spoken language processing applications including discourse analysis, turn taking or speaker interaction pattern analysis. With the increasing amount of dyadic or small group interaction data being generated requiring analysis, manual marking of speech segments becomes equally expensive and time consuming. There have been previous efforts in automatically detecting speech segments using multiple-state hidden markov model (HMM), where states either represent 'speech' and 'nonspeech' [1, 2] or correspond to 'speech', 'overlapped speech', 'crosstalk', and 'silence' [3]. One of the frequently faced problems is that the recorded signal in each channel of a multi-speaker (e.g., meeting) corpus often contains crosstalk [4] – the occurrence of the signal in the channel from sources other than the intended primary source for that channel. There can also be time segments where simultaneous voices from two or more participants get recorded resulting in overlapped speech. Modeling these various events using HMM states requires manually tagged training data and is also vulnerable to channel variation.

Laskowski et. al. [5] proposed an efficient cross-correlation-based voice activity detection (VAD) scheme, which requires no prior training and executes in one fifth real time. Fast multichannel VAD schemes without prior training yet robust to noise or

channel variation are desirable since most of the data processing and analysis following VAD such as speech recognition are generally more time-consuming. In this paper, we consider only stereo recording and propose a simple but robust VAD scheme for two channel (stereo) speech signal with crosstalk. The proposed scheme consists of a standard VAD on the merged signal from two channels followed by a channel selection strategy. The goal of the channel selection is to determine the target channel at every short-time frame when voice activity is detected in the merged signal. There is no need for prior training in the proposed approach. Since microphones are close-talking, the channel corresponding to the target speaker's microphone will have higher energy compared to the same speech from other channel. Thus energy of the speech signal in different channels could be a potential indicator for selecting the target channel during speech. In addition to this energy based channel selection, we found that the acoustic characteristics of the merged signal are closer to that of the primary channel signal compared to that of the cross-talk channel signal. We also used this property to perform channel selection. We found that a long-term analysis window around the target frames improves the channel selection and overall VAD performance. Experiments with recordings of real conversations from cross-lingual doctor patient interactions show that the VAD performance using the proposed method is further improved compared to the cross-correlation-based VAD proposed in [5].

We begin with the description of the dataset (section 2) used in our experiment. The proposed approach for joint VAD and channel selection is described in section 3. Experiments and results of VAD on stereo recording is explained in section 4. Conclusions are drawn in section 5.

2. Data: Cross Lingual Medical Interactions

As a part of the medical domain speech-to-speech (s2s) translation project, we have recorded conversations between an English doctor and a Spanish patient with an interpreter in between. The recording was done in a typical office room setting with some background noise from the air-conditioner. The doctors are students from USC's Keck School of Medicine while the Spanish speaking patients are standardized patients (actors trained to assess Doctor skills). Three close-talking microphones were used for the doctor, patient, and interpreter. The Microtrack II professional 2-channel mobile digital recorder was used for stereo recording with sampling frequency 48kHz. The microphones of the doctor and the patient were connected to the left channel of the recorder and the interpreter to the right. Since the doctors and patients do not know each others' language, they communicated through the interpreter and have practically no overlap in their speech activity; therefore, the doctor and the patient are recorded on the same channel. From our entire collection, we have used
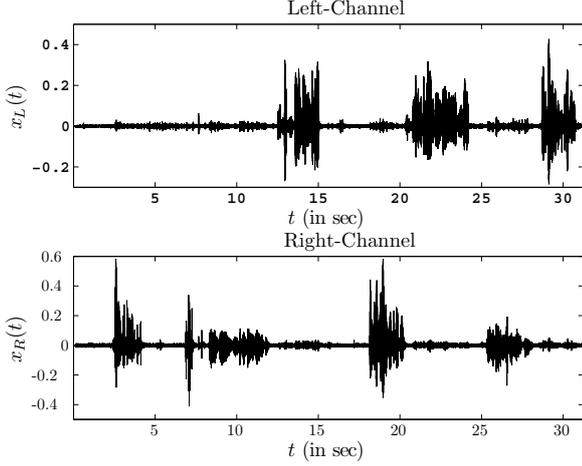
Figure 1: Sample signal shows a significant crosstalk. Right channel contains crosstalk over the following durations [12.5, 15.5], [20.5, 25], and [27, 31] sec. Similarly, crosstalks in left channel happen during [2.5, 12.5], [17.5, 20], and [25, 27] sec.

5 sessions of recordings for our experiments with an average duration of 30 minutes per session. During the collection, the patient and interpreter were sitting side by side while the doctor was sitting across approximately 1 meter away. This proximity and alignment of the participants resulted in a significant crosstalk. A sample recording for ~30 seconds is illustrated in Fig. 1. The crosstalk in both channels because of the speech from the participants in other channels is clearly visible.

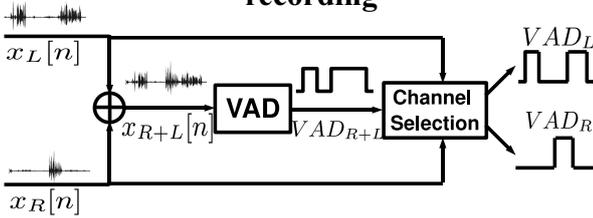## 3. The proposed approach of VAD in stereo recording

Figure 2: Schematic diagram of the VAD for stereo recording.

The proposed joint VAD and channel selection scheme is shown using a block diagram in Fig. 2. Let $x_R[n]$ and $x_L[n]$ denote the signal recorded in the right and the left channels respectively. Let us define $x_{R+L}[n] = x_R[n] + x_L[n]$. $x_{R+L}[n]$ is a single channel signal; hence, the standard VAD schemes can be used to detect the presence of speech in the merged signal at each frame (frame index m) with frame duration of N_w samples and frame shift of N_{sh} samples. The presence of speech in $x_{R+L}[n]$ at any frame indicates that there is speech either in the right or the left channel in the respective frame. In other words, when $\text{VAD}_{R+L}(m)=1$, we need to determine whether $\text{VAD}_R(m)=1$ or $\text{VAD}_L(m)=1$. This is done using the channel selection scheme which takes $\text{VAD}_{R+L}(m)$ and signals from two channels over a range of frame indices $[m - R, \ m + R]$ as inputs and produces the VAD for each channel (VAD_R and VAD_L). The purpose of performing VAD before channel selection is to determine the time segments when there is speech from the participants in the conversation. Unlike [5], we make a realistic assumption that the speech need not be present at every time instant and hence the channel selection should be performed only when there is speech in one of the channels. Below, we describe the channel selection scheme.

Suppose there are N samples observed from both right and left channels, i.e., $x_R[n]$ and $x_L[n], 0 \leq n \leq N - 1$. Using the standard VAD on the combined signal, we know that there is speech from a speaker of the left or the right channel during the observation. Without loss of generality, let us assume that the speaker with the close-talking microphone corresponding to the right channel is speaking during this observation which implies that there is a leakage from the right channel to the left channel. Thus, we can write the observed signals in the two channels as follows:

$$\left. \begin{array}{l} x_R[n] = S[n] + N_1[n] \\ x_L[n] = \alpha S[n - n_0] + N_2[n] \end{array} \right\} 0 \leq n \leq N - 1 \qquad (1)$$

where $N_1[n]$ and $N_2[n]$ are additive white noises and are assumed to be not only independent of each other but also independent of $S[n]$. We assume that $N_1[n]$ and $N_2[n]$ have zero mean and variance σ^2, $\forall n$. $S[n]$ is the speech of the participant in the right channel. $\alpha S[n - n_0]$ is the leakage in the left channel, where α denotes the level of leakage of the speech from the right to the left channel and n_0 denotes the delay in number of samples between the $S[n]$ and its leaked version. We also assume that $0 < \alpha < 1$ based on the longer distance of the active speaker from the left microphone and the directionality of the close-talk microphones.

3.1. Energy-based channel selection

The power of the signal in the right channel ($\sigma_{x_R}^2$) is more compared to that of the left channel ($\sigma_{x_L}^2$). Let σ_s^2 be the variance of the speech signal (assuming $S[n]$ is zero mean). Then,

$$\begin{aligned} \sigma_{x_R}^2 &= \sigma_s^2 + \sigma^2 \\ &> \alpha^2 \sigma_s^2 + \sigma^2 = \sigma_{x_L}^2 (\because 0 < \alpha < 1) \\ \implies \sigma_{x_R}^2 &> \sigma_{x_L}^2 \end{aligned} \qquad (2)$$

Thus, if the power of the right channel is more than that in the left channel (i.e., $\sigma_{x_R}^2 > \sigma_{x_L}^2$), then the target channel at the respective frame is the right channel, i.e., $\text{VAD}_R(m)=1$. We refer this energy based channel selection scheme by ENERGY.

Let $\mathbf{F}_l^R(\omega) = |X_l^R(\omega)|^2$, $\mathbf{F}_l^L(\omega) = |X_l^L(\omega)|^2$, and $\mathbf{F}_l^{R+L}(\omega) = |X_l^{R+L}(\omega)|^2$ denote the Fourier spectra of $x_R[n]$, $x_L[n]$, and $x_{R+L}[n]$ at the lth frame. Note that $X_l^{R+L}(\omega) = X_l^R(\omega) + X_l^L(\omega)$, where ω denotes the angular frequency. Considering non-overlapping frames, we can write eqn. (2) using Parseval's theorem as follows:

$$\sum_{l=m-R}^{m+R} \int_{-\pi}^{\pi} \mathbf{F}_l^R(\omega) d\omega > \sum_{l=m-R}^{m+R} \int_{-\pi}^{\pi} \mathbf{F}_l^L(\omega) d\omega \qquad (3)$$

$$\implies \sum_{l=m-R}^{m+R} \int_{-\pi}^{\pi} \left| X_l^{R+L}(\omega) - X_l^L(\omega) \right|^2 d\omega$$

$$> \sum_{l=m-R}^{m+R} \int_{-\pi}^{\pi} \left| X_l^{R+L}(\omega) - X_l^R(\omega) \right|^2 d\omega$$

This means that the short-time spectra of the combined signal will be closer to that of the primary channel as opposed to the crosstalk channel. This motivates us to explore the perceptual distance between short-time spectra and other spectral feature based on acoustic proximity measures for the task of channel selection.

3.2. Acoustic proximity measures for the channel selection

We have used two measures of acoustic and perceptual proximity (discussed below) between two signal segments to determine whether $x_{R+L}[n]$ is closer to $x_R[n]$ or $x_L[n]$.

3.2.1. Euclidean distance between MFCC features

The mel-frequency cepstral coefficients (MFCC) [6] capture the energy distribution over different frequency bands for a short-time speech segment. Let $\mathbf{c}_l^R$, $\mathbf{c}_l^L$, and $\mathbf{c}_l^{R+L}$ denote the MFCC vectors of $x_R[n]$, $x_L[n]$, and $x_{R+L}[n]$ respectively at the lth frame. The distance between MFCC of $x_{R+L}[n]$ and $x_R[n]$ over the $2R+1$ observed frames is computed as

$$\delta_{R+L,R} = \sum_{l=m-R}^{m+R} ||\mathbf{c}_l^{R+L} - \mathbf{c}_l^R||_2^2 \qquad (4)$$

where $|| \cdot ||_2$ denotes the L_2 norm of a vector. Similarly, $\delta_{R+L,L}$ is computed. If $\delta_{R+L,R} < \delta_{R+L,L}$, it indicates that the spectral characteristics of $x_{R+L}[n]$ are closer to that of $x_R[n]$ than that of $x_L[n]$ over the observed signal segment. Thus, when $\delta_{R+L,R} < \delta_{R+L,L}$, $\text{VAD}_R(m)$ is set to 1, otherwise $\text{VAD}_R(m)$ is set to zero. Decision for VAD_L is done in a similar way. We refer to this channel selection scheme by MFCC_0. We also consider a channel selection scheme without the zero-th coefficient of MFCC, which we refer only by MFCC.

3.2.2. Itakura-Saito distance

Itakura-Saito (IS) distance [7] is a measure of the perceptual difference between the spectra of two short-time signal segments. The distance between $x_{R+L}[n]$ and $x_R[n]$ over the $2R+1$ observed frames using frame-based IS distance is computed as

$$\delta_{R+L,R} = \sum_{l=m-R}^{m+R} d\left(\mathbf{F}_l^R(\omega), \mathbf{F}_l^{R+L}(\omega) \right) \qquad (5)$$

where, $d\left(\mathbf{F}_l^R(\omega), \mathbf{F}_l^{R+L}(\omega) \right)$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{\mathbf{F}_l^R(\omega)}{\mathbf{F}_l^{R+L}(\omega)} - \log \frac{\mathbf{F}_l^R(\omega)}{\mathbf{F}_l^{R+L}(\omega)} - 1 \right] d\omega$$

$\delta_{R+L,L}$ is computed in a similar way. Similar to the description in section 3.2.1, $\text{VAD}_R(m)$ is set to 1 when $\delta_{R+L,R} < \delta_{R+L,L}$, otherwise $\text{VAD}_R(m)$ is set to zero. VAD_L is determined similarly. We refer to the IS based channel selection scheme by IS.

Note that, under the signal model as defined in eqn. (1), the energy-based channel selection scheme (eqn. (2)) is independent of the noise variance σ^2. However, this may not be true for MFCC, MFCC_0 and IS channel selection schemes. Also, in practice, noise variance in both channels need not be identical; in such cases, MFCC, MFCC_0, and IS could be more robust compared to energy-based channel seleccction scheme. Due to the non-linear function involved in the MFCC feature computation and IS distance computation, they are not analytically tractable unlike the energy-based approach. Also, in practice, the two channel signal model defined in eqn. (1) is a simplification (e.g. it doesn't consider reverberation) and we would like to explore the utility of MFCC, MFCC_0 and IS for such cases through experiments.

4. Experimental Results

We manually labeled the speech segments in 5 sessions of stereo recording (overall $\sim$2.5 hours of two channel audio) to obtain reference VAD decisions which are used to evaluate the proposed approach. As a baseline, we chose the VAD decisions in each channel obtained by the ETSI AMR VADs option 2 [8] (AMRVAD2). The implementation was taken from [9]. We have downsampled the speech signal in both channels to 8kHz for VAD so that the sampling frequency satisfies the AMRVAD2 requirements. Based on the analysis in section 3, we used the VAD on the combined
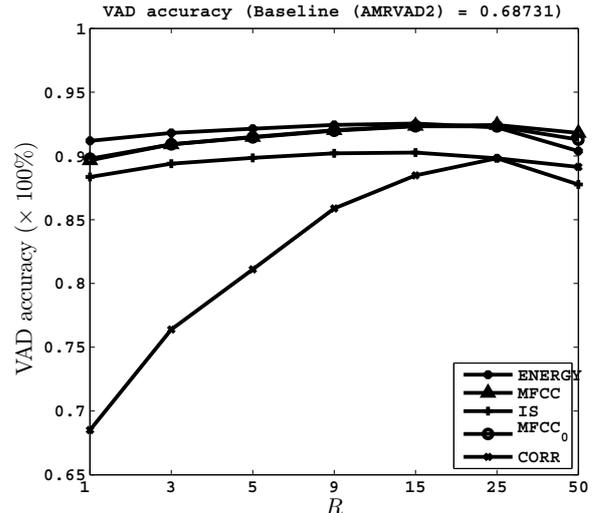

Figure 3: Stereo VAD accuracy obtained by different methods for varying R compared against the baseline accuracy of 68.73%.

signal, $x_{R+L}[n]$, followed by ENERGY, MFCC, IS, and MFCC_0 schemes to obtain channel specific VAD decisions. For comparison, we also used the cross-correlation based channel selection scheme, similar to [5], which is denoted by CORR. This is done by finding the peak in the cross-correlation between signals of two channels and determining the delay between them.

VAD decisions are made every 10 msec (N_{sh}=80 samples) and the short-time frame duration is chosen to be 20 msec (N_w=160 samples). As described in section 3, we consider the signal segment over the frame indices $[m-R,\ m+R]$ to determine $\text{VAD}_R(m)$ and $\text{VAD}_L(m)$. We experiment by varying the value of R from the following set $\{1, 3, 5, 9, 15, 25, 50\}$; R=1 corresponds to the case when no signal from neighboring frame is considered and R=50 corresponds to the case when a signal segment of approximately 1 sec is considered with the current frame in the middle. Also, note that channel selection scheme is applied only to those frames for which VAD_{R+L}=1.

We plot the VAD accuracy obtained by different channel selection schemes for varying R in Fig. 3. VAD accuracy is determined by the number of correctly detected speech and non-speech frames against the reference VAD decisions. It is clear that the VAD accuracies obtained using ENERGY, MFCC, MFCC_0 are similar and higher than those obtained by IS and CORR. Fig. 3 also indicates that the correlation based channel selection scheme is not realiable particularly over short frame lengths. This could be due to the fact that over short frame length a significant peak may not be observed in the cross-correlation estimate due to background noise; over long frame length, the estimate of the cross-correlation would be better and hence the effect of noise will be reduced. We observe that the VAD accuracy increases as R increases consistently for all the different channel selection schemes; when R goes above 15 or 25, the VAD accuracy starts to decrease. The highest VAD accuracy of 92.54% is obtained by ENERGY for R=15, which is $\sim$24% absolute improvement over baseline. The best VAD accuracies obtained by MFCC and MFCC_0 are 92.43% and 92.33% respectively for R=25 and R=15. This means that although ENERGY and MFCC do not exploit identical acoustic properties of the signal, they yield similar VAD performance. Note that the VAD accuracy in individual channels obtained as a result of the channel selection scheme is affected by the VAD on the combined signal. The VAD accuracy on the combined signal using AMRVAD2 is $\sim$90%. Since channel
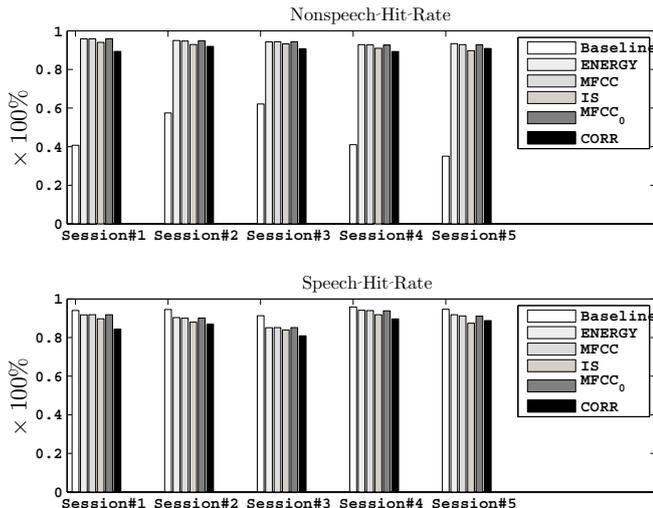
Figure 4: Speech and non-speech hit rate obtained by different channel selection schemes for five sessions considered at R=15.

Figure 5: Different VAD performance evaluation measures obtained by the proposed schemes at R=15.

selection is performed only in the frames where VAD_{R+L}=1, incorrect detection of a speech frame will affect the performance of VAD in both channels; but mis-detection of a non-speech frame will affect the performance of only one channel. A better VAD algorithm on the combined signal can result better VAD performance for the individual channels after channel selection.

In Fig. 4, we plot the speech hit rates (i.e., number of correctly detected speech frames) and non-speech hit rates for each session obtained by different channel selection schemes at R=15. It is clear that the performance of different schemes is consistent across sessions. It is also clear that the baseline non-speech hit rate is significantly poor compared to the proposed schemes including CORR for all sessions. Thus, the overall VAD accuracy improves because of the improvement in non-speech hit rate. In other words, the effect of crosstalk on VAD is significantly reduced by the proposed scheme. For a comprehensive evaluation of the proposed stereo VAD schemes, we follow the testing strategy proposed by Freeman et al [10], where five different parameters reflecting the VAD performance are considered: (1) CORRECT, (2) FEC (front end clipping), (3) MSC (mid speech clipping), (4) OVER (carry over), and (5) NDS (noise detected as speech). Fig. 5 shows these five parameters obtained by various VAD schemes at R=15 over all five recording sessions. We observe that five evaluation parameters obtained by ENERGY, MFCC, $MFCC_0$ are similar to each other and are consistently better than the baseline and of those obtained by IS and CORR indicating the effectiveness of the proposed VAD scheme in the presence of crosstalk.

5. Conclusions

From the results of the VAD experiments, we find that consideration of a long-term analysis window yields better results compared to short-time frame. This is consistent with the findings in [11]. A long-term approach based VAD could be used on the combined signal as well to improve the stereo VAD accuracy further. It should be noted that the proposed joint VAD and channel selection scheme can be easily extended to apply for audio recordings with more than two channels. We obtain ∼8% VAD error over 2.5 hours of stereo audio considered in our experiment. One potential reason for the remaining errors could be the overlapping speech in the two channels, which is not considered in
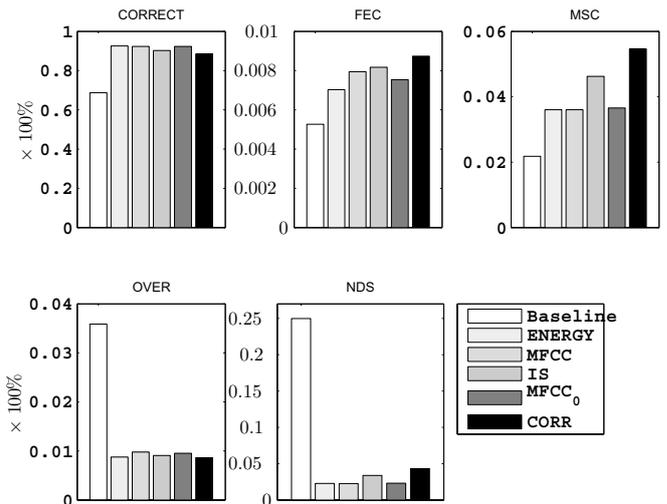
the proposed two channel signal model. Although the overlapping speech is a small fraction in the stereo recording considered in this paper, in practice it could be significant portion of any natural interaction. Thus, considering overlapped speech within the framework may improve the VAD accuracy and is part of our future work.

6. References

[1] T. Pfau and D. P. W. Ellis, "Hidden markov model based speech activity detection for the ICSI meeting project", *Proc. Eurospeech*, Aalborg, Denmark, September, 2001.

[2] T. Pfau, D. Ellis, and A. Stolcke, "Multispeaker Speech Activity Detection for the ICSI Meeting Recorder", *Proc. ASRU*, Trento, Italy, pp. 107-110, Dec 2001.

[3] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multi-channel audio", *IEEE Trans. Speech and Audio Proc.*, vol. 3, issue 1, pp. 84-91, Jan 2005.

[4] H. M. Chang, "CrossTalk: technical challenge to VAD-like applications in mixed landline and mobile environments", *Proc. 3rd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, pp. 77-80, Oct 1996.

[5] K. Laskowski, Q. Jin, and T. Schultz, "Cross-correlation-based multispeaker speech activity detection", *Proc. Interspeech*, ICSLP, Jeju Isand, Korea, pp. 973-976, Oct 2004.

[6] M. Xu, L. Y. Duan, J. Cai, L. T. Chia, C. Xu, and Q. Tian, "HMM-based audio keyword generation", *Kiyoharu Aizawa, Yuichi Nakamura, Shin'ichi Satoh. Advances in Multimedia Information Processing - PCM 2004: 5th Pacific Rim Conference on Multimedia. Springer*.

[7] R. Gray, A. Buzo, A. Gray, and Y. Matsuyama, "Distortion measures for speech processing", *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. 28, no. 4, pp. 367-376, Aug 1980.

[8] *Digital Cellular Telecommunications System (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi Rate (AMR) Speech Traffic Channel; General Description*, 1999.

[9] *Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi Rate (AMR) Speech; ANSI-C code for AMR Speech Codec*, 1998.

[10] Freeman D. K., Southcott C. B., Boyd I., and Cosier G., "A voice activity detector for pan-European digital cellular mobile telephone service", *Proc. ICASSP*, Glasgow, U.K., vol. 1, pp 369-372, 1989.

[11] Ramirez J., Segura J. C., Benitez C., Torre A., and Rubio A., "Efficient voice activity detection algorithms using long-term speech information", *Speech Communication*, vol. 42, issues 3-4, pp 271-287, April 2004.