Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 930376, 8 pages
doi:10.1155/2010/930376

Research Article

Multimodal Speaker Verification Based on Electroglottograph Signal and Glottal Activity Detection

Zoran Ćirović,[1] Milan Milosavljević,[2,3] and Zoran Banjac[1]

[1] *School of Electrical Engineering and Computer Science, Vojvode Stepe 283, 11000 Belgrade, Serbia*
[2] *Faculty of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, 11000 Belgrade, Serbia*
[3] *University Singidunum, Danijelova 29, 11000 Belgrade, Serbia*

Correspondence should be addressed to Milan Milosavljević, mmilan@etf.rs

Received 26 March 2010; Revised 2 August 2010; Accepted 28 August 2010

Academic Editor: Sharon Gannot

To achieve robust speaker verification, we propose a multimodal method which includes additional nonaudio features and glottal activity detector. As a nonaudio sensor an electroglottograph (EGG) is applied. Parameters of EGG signal are used to augment conventional audio feature vector. Algorithm for EGG parameterization is based on the shape of the idealized waveform and glottal activity detector. We compare our algorithm with conventional one in the term of verification accuracy in high noise environment. All experiments are performed using Gaussian Mixture Model recognition system. Obtained results show a significant improvement of the text-independent speaker verification in high noise environment and opportunity for further improvements in this area.

1. Introduction

Speaker Verification (SV) is the process of verifying the claimed identity of a speaker using features extracted from her/his voice. Conventional SV uses the recorded audio signal as the sole source of information. This is based on features such as linear predictive cepstral coefficients (LPCC), mel-frequency cepstral coefficients (MFCC), or log area ratio (LAR) [1–3]. Over the past several years, one of the dominant approaches for modeling in text-independent SV applications has been based on Gaussian mixture models (GMMs) [1, 4–7].

In the case of speech being corrupted by environmental noise, the distribution of the audio feature vectors is also damaged. This leads to misclassification and poor recognition. For an SV system to be of practical use in a high noise environment it is necessary to address the issue of robustness. To combat this problem, researchers have put forward several new algorithms, which assume prior knowledge of the noise, like noise filtering techniques [8, 9], parallel model combination [10–12], Jacobian environmental adaptation [13, 14], using microphone arrays [15, 16], or techniques of speech enhancement which target the modeling of speech and noise pdf [17, 18]. When there is insufficient knowledge of the noise, one may attempt to ignore the contribution of highly corrupted speech data [19, 20] or to combine multicondition model training and the missing-feature theory to model noise with unknown temporal-spectral characteristics [21].

It is possible to accomplish robustness by the utilization of other sensing modalities to complement the audio signal of speech. As a matter of fact, in almost every context, carefully designed multimodal interfaces turned out to be more beneficial than any single-modality interface [1, 22]. Some multimodal approaches are based on sensors where a speaker is not connected to a recording device, like GEMS, ultrasonic or video signal [23–25]. Other researches use sensors physically connected to the speaker's head, face or throat, like electroglottograph (EGG), P-microphone, bone-conducting microphone [22, 24, 26]. The practical application of physically connected sensors is in specific environment (military approach, battle field environment, etc.) as well as in situations where the user is willing to cooperate meaning amenable to attach the sensor on herself/himself.

This study demonstrates that the specificity of the EGG waveform is different relative to different speakers (see

Section 3). We use EGG features representing the time characteristics of an idealized EGG waveform. Then, we concatenate both the EGG features and audio features by applying a glottal activity detector. The main contribution of this paper is to investigate the performance of this fusion for SV problem in a high noise environment. In this research, we also discuss the selection of an activity detector.

There are two stages in the SV process (see Figure 1). The first is enrollment (training), where model parameters $\boldsymbol{\lambda}_i$ are computed for each registered speaker,"i", $i = 1, 2, ..., N$, represented by the feature vectors $\mathbf{X}_i$. In the proposed SV, $\mathbf{X}_i$ presents the new feature vectors, given by

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{X}_{i,\mathrm{MFCC}} \\ \mathbf{X}_{i,\mathrm{EGG}} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^1_{i,\mathrm{MFCC}} & \mathbf{x}^2_{i,\mathrm{MFCC}} & \cdots & \mathbf{x}^{L_i}_{i,\mathrm{MFCC}} \\ \mathbf{x}^1_{i,\mathrm{EGG}} & \mathbf{x}^2_{i,\mathrm{EGG}} & \cdots & \mathbf{x}^{L_i}_{i,\mathrm{EGG}} \end{bmatrix}, \quad (1)$$

where L_i is number of feature vectors for ith speaker; $\mathbf{X}_{i,\mathrm{MFCC}}$ and $\mathbf{X}_{i,\mathrm{EGG}}$ are the sequences of the audio and EGG feature vectors $\mathbf{x}^k_{i,\mathrm{MFCC}}$, $\mathbf{x}^k_{i,\mathrm{EGG}}$, respectively, where $k = 1, ..., L_i$.

During this stage, the background model $\boldsymbol{\lambda}_{\bar{i}}$ (alternative hypothesis model) is created for each speaker using $N - 1$ vectors $\mathbf{X}_j$, $j = 1, ..., N$; $j \neq i$, which do not belong to a certain person "i". The background model becomes invariant and common for to all $\mathbf{X}_i$ if $N \to \infty$. In the second (testing) stage, the classifier decides whether the new input utterance, denoted by $\mathbf{X}_{\mathrm{test}}$, belongs or not to the claimed registered speaker, represented by model $\boldsymbol{\lambda}_{\mathrm{claim}} \in \{\boldsymbol{\lambda}_i\}$, $i = 1, ..., N$, by comparing the conditional probabilities $P(\mathbf{X}_{\mathrm{test}}/\boldsymbol{\lambda}_{\mathrm{claim}})$ versus $P(\mathbf{X}_{\mathrm{test}}/\boldsymbol{\lambda}_{\overline{\mathrm{claim}}})$, where $\boldsymbol{\lambda}_{\overline{\mathrm{claim}}}$ corresponds to the background model, [5].

Computing of feature vectors (parameterization) is common to both stages. The different nature of audio and EGG signals requires specific methods for optimal parameterization.

2. Parameterization

Parameterization is the transformation of an input signal into a set of feature vectors which are less redundant and more suitable for statistical modeling than the input signal. The input signal is processed into frames creating a sequence of vectors. Each frame corresponds to a time window t_W with overlapping between the consecutive frames.

The multimodal speaker verification, proposed in this work, includes audio and EGG parameterization.

2.1. Audio Parameterization. Audio parameterization is usually based on the cepstral representation of an audio signal, [6]. Prior to computing a short-term power spectra, the audio signal is filtered with a first-order FIR filter to spectrally flatten the signal. Pure cepstral coefficients of a speaker "i", denoted by $\mathbf{X}_{i,C}$, are obtained applying the mel-scaled filter banks up to 4 kHz. Time derivatives of cepstral coefficients are resistant to linear channel mismatches between training and testing and have yielded significant improvement in the recognition processes, [27]. These coefficients $\Delta\mathbf{X}_{i,C}, \Delta\Delta\mathbf{X}_{i,C}$ are derivatives of the time function of the cepstral coefficients and are, respectively, called the delta- and delta-delta-cepstral coefficients. Regarding this, vector $\mathbf{X}_{i,\mathrm{MFCC}}$ is

$$\mathbf{X}_{i,\mathrm{MFCC}} = \begin{bmatrix} \mathbf{X}_{i,C} \\ \Delta\mathbf{X}_{i,C} \\ \Delta\Delta\mathbf{X}_{i,C} \end{bmatrix}. \quad (2)$$

2.2. EGG Parameterization. The electroglottograph is a device for the measurement of the time variation of the degree of contact between vibrating vocal folds during voice production. The degree of contact is proportional to the impendence between two electrodes on the subject's neck when the current is in the MHz region. Typical waveforms of EGG and related audio signal are shown on Figure 2.

For unvoiced segments, the EGG waveform contains slow changes and very low-level high-frequency noise that is easily distinguished, [28]. To remove disturbing low-frequency (uninformative) fluctuations, the EGG signal is usually filtered, using digital linear phase high or bandpass filters.

The EGG signal can be considered as "almost periodic" in voiced segments. One period of EGG signal with characteristic segments is shown in Figure 3, synchronously with audio signal.

For voiced segments, the EGG usually has only two zero crossings per fundamental (pitch) period of voicing. In order to obtain a quantitative description of the EGG signal, a model based on the shape of the idealized waveform as proposed in [29, 30] is used. The idealized waveform has flat characteristics intervals although the original signal has a typically parabolic shape. When the vocal folds are open and it is ensured that there is no lateral contact between the vocal folds, the impedance is maximal and peak glottal flow occurs—*open phase*. The EGG waveform in this segment is flat, with small fluctuations. Further on, the movements of the margins of the vocal folds come into the contact and the vocal folds continue to close—*closing phase*. During the closing phase the vocal folds remain in contact and the airflow is blocked. Like in the open phase, limited fluctuations of the impedance are observed. However, the waveform is not flat, but rather forms a smooth hill. Pitch period—T_0 and specific durations in EGG waveform: t_1, t_2, t_3, t_4 are marked in Figure 3. Time from the maximum contact to zero crossing (about half of the opening phase) is marked as t_1. Time t_2 is next interval up to the maximum of the open phase. Considering that the open phase is rather flat, t_2 could be calculated to the mean of the open phase of idealized waveform. Next, t_3, t_4 are intervals to second zero crossing and to the maximum in the contact phase, respectively.

Assuming that the EGG signal contains specific information about the speaker (see Section 3) and that EGG sensor is robust in noisy environments [22], adding related parameters to the features in the SV process, is expected to be beneficial. The EGG features used are period of the fundamental frequency T_0 and a set of timing parameters:

$$\mathbf{x}^n_{\mathrm{EGG}} = \left[T_0 \;\; \overline{t_1} \;\; \overline{t_2} \;\; \overline{t_3} \;\; \overline{t_4} \right]^T, \quad (3)$$

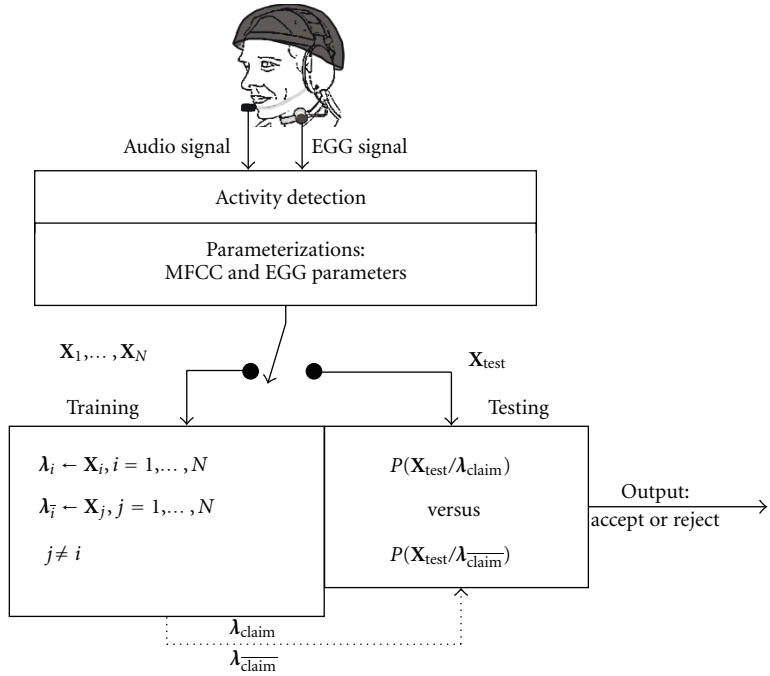


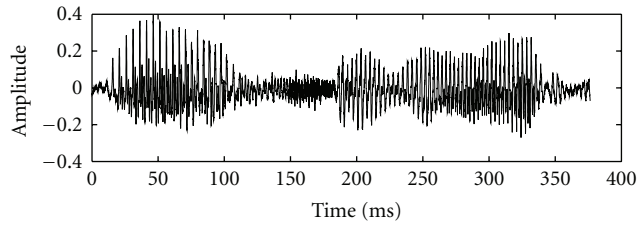
Figure 1: Stages of the SV process: parameterization (with activity detection), training, and testing.

where $\overline{t_1}$ to $\overline{t_4}$ are normalized time parameters t_1 to t_4, with respect to T_0, measured at the time instant n. These features are correlated with the most salient glottal phenomena, that is, glottal pulse width, skewness, and abruptness of closure.

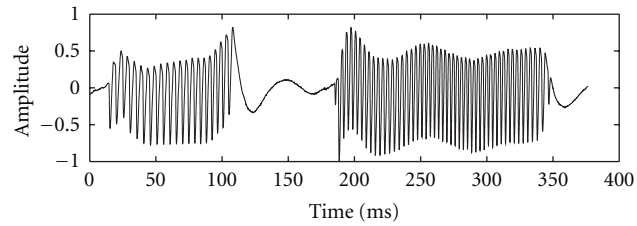
Natural speech consists of speech segments, silence, and background noise. To only extract features from speech segments, the input signals are first fed to the activity detector subsystem to separate speech from nonspeech. Based on the activity detector output, features are extracted and then normalized.

2.3. Activity Detection. Voice activity detector (VAD) is a preprocessing subsystem designed for distinguishing speech from nonspeech segments in an audio signal. Conventional VAD algorithm is based on energy and zero-crossing rate or cepstrum [31]. In a multimodal system, distinguishing speech could be based on the additional signal produced by nonaudio sensors.

While EGG signal retains the relevant information on the excitation source, for only the voiced segments of the speech signal, classical VAD includes both voiced and unvoiced, the glottal activity detector (GAD) is used for the EGG features extraction and fusion with cepstral coefficients in the multimodal feature vectors.



(a) Audio signal



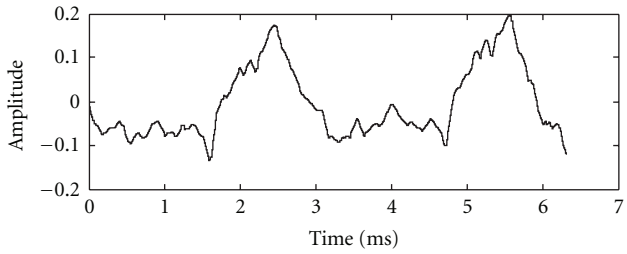
(b) EGG signal

Figure 2: A speech segment represented by: (a) audio and (b) EGG waveform.

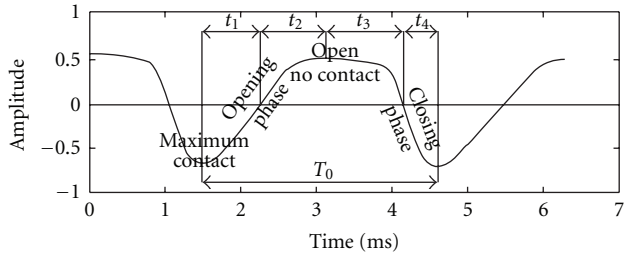
3. Expected Discrimination Information of EGG Features

Discrimination property of EGG features in the proposed SV system could be analyzed and estimated in two ways: (i) *a priori*, without the design of the classification system and system's performance estimation; (ii) *a posteriori*, comparing the accuracy of the verification systems with or without augmented EGG features. *A posteriori* approach will be considered in Section 4, relative to the accuracy of the analysis of the proposed system.

Based on the fact that the proposed SV system compares the probabilities of GMM models, for the approach (i), discrimination property of EGG features can be measured by using Kullback-Leibler divergence (KLD) between corresponding probability distributions of EGG features, [32].



(a) Audio signal



(b) EGG signal

Figure 3: One pitch period of (a) audio and (b) EGG waveform, showing the glottal opening and closure phases.

For each speaker the GMM model $\boldsymbol{\lambda}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \mathbf{w}_i\}$, $i = 1, \ldots, N$ is created as mixture of M Gaussian densities

$$p(\mathbf{x}^n_{\text{EGG}}/\boldsymbol{\lambda}_i) = \sum_{m=1}^{M} w_m g(\mathbf{x}^n_{\text{EGG}}/\boldsymbol{\lambda}_i) \,, \tag{4}$$

where $\mathbf{x}^n_{\text{EGG}}$ is an nth EGG feature vector, as in (3); w_m represents weights, where $\sum_{m=1}^{M} \mathrm{w}_m = 1$ and $\mathrm{g}(\mathbf{x}^n_{\text{EGG}}/\boldsymbol{\lambda}_i)$ is single Gaussian density with mean vector $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{\Sigma}_m$.

KLD is the fundamental measure between the statistical distributions, which quantifies how close a probability distribution $p(x)$ is to another distribution $q(x)$

$$\text{KLD}(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \,. \tag{5}$$

$\text{KLD}(p(\mathbf{x}_n/H)\|p(\mathbf{x}_n/\overline{H}))$ can be interpreted as the expected discrimination information between the null and alternative statistical hypotheses, for discriminating in favor for a hypothesis H, against hypothesis $\overline{H}$, when hypothesis H is true. If H represents a model denoted by $\boldsymbol{\lambda}_i$, which characterizes the hypothesized speaker in the feature space of $\mathbf{x}^n_{EGG}$, and $\overline{H}$ represents another model $\boldsymbol{\lambda}_j$, $j \neq i$, expected discrimination information becomes

$$\text{KLD}(i\|j) = \text{KLD}\left(p(\mathbf{x}^n_{\text{EGG}}/\boldsymbol{\lambda}_i)\|p\left(\mathbf{x}^n_{\text{EGG}}/\boldsymbol{\lambda}_j\right)\right), \quad j \neq i, \tag{6}$$

where $\boldsymbol{\lambda}_i, \boldsymbol{\lambda}_j$ are only based on corresponding EGG features.

$\text{KLD}(i\|j)$, defined in (6) is measured, when models $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_j$: (a) belong to the same speaker, denoted by $\text{KLD}_{\text{intra}}(i\|j)$ (intraspeaker variability) and (b) belong to the different speakers, denoted by $\text{KLD}_{\text{inter}}(i\|j)$ (interspeaker variability). Figure 4 presents results for six speakers in the form of histogram.

Table 1: Experiments for analyzing EGG signal contribution in high noise environment.

	Vector features	Activity Detector
System 1	$\mathbf{X}_{\text{MFCC}}$	VAD
System 2	$\mathbf{X}_{\text{MFCC}}$	GAD
System 3	$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{\text{MFCC}} \\ \mathbf{X}_{\text{EGG}} \end{bmatrix}$	GAD
System 4	$\mathbf{X}_{\text{EGG}}$	GAD

From Figure 4, it is clear that there is almost no overlapping between the two groups of divergences $\text{KLD}_{\text{intra}}(i\|j)$ and $\text{KLD}_{\text{inter}}(i\|j)$. For all speakers in database (see Section 4.1), $\max(\text{KLD}_{\text{intra}}(i\|j)) < KLD_{\text{inter}}(i\|j)$ is true in 94.2%. Considering this result, one can conclude that EGG features, have speaker discriminative property, but the contribution of these features, in the process of speaker verification, will be examined in the following experiments.

4. Experiments

This section analyzed contribution of EGG features to the proposed system, when the audio signal has been corrupted by additive White-Gaussian noise. The proposed SV system is compared to the audio-based (conventional) SV system. In order to clearly show the contribution of EGG features, four experiments were conducted. The first experiment was conducted by using the conventional system and conventional VAD. The second experiment was identical to the first, except that VAD was replaced by GAD. The third experiment involves multimodal parameters (audio and EGG features) with the addition of GAD. The fourth experiment involves only EGG features and GAD. SV error rates (ERRs) for different values of SNR (from 0 to 30 dB) were analyzed for all experiments. These experiments are illustrated in the four systems summarized in Table 1.

4.1. Database. The corpus consists of 50 sessions with 16 speakers with up to 4 sessions per speaker. The utterances for each session were very carefully chosen to provide a very good representation of typical Serbian language [33]. Audio and EGG signals were recorded by microphone and an EGG device (model EG-PC3 produced by Tiger DRS, Inc., USA) synchronously. Both signals were originally sampled at 44 kHz. We used one session as enrollment and the remaining 49 sessions were used for speaker verification. This resulted in $49 * 50 = 2450$ speaker verifications tests.

4.2. Conventional Verification System (System 1). Conventional verification system consists of the front-end audio signal processing in order to produce feature vector, as in (2). The audio feature vector is formed as a collection of 14 mel-frequency cepstral coefficients, plus corresponding deltas, altogether $D = 42$ coefficients per frame. Each frame corresponds to 1024 samples, for example, $t_W \cong 23.2$ ms time window. The frames are overlapped to avoid the risk of losing valuable transient. In our system, frames are overlapped by one half of the frame length. After computing

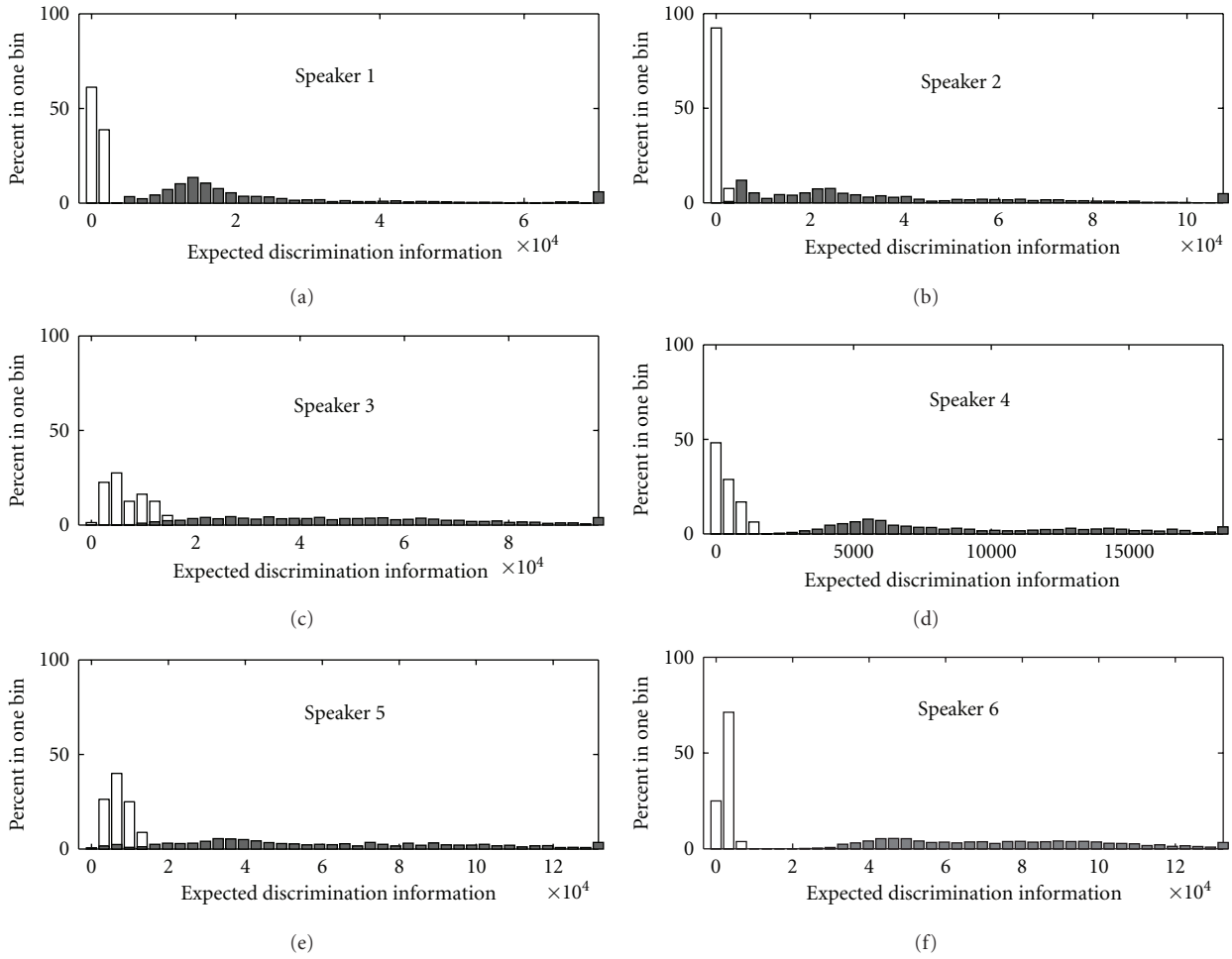


Figure 4: Expected discrimination information for six speakers, when models belong to the same speaker (white bars) and belong to the different speakers (gray bar).

the MFCCs, a cepstral mean subtraction was done, [5]. To separate speech frames from silent and noise, classical VAD based on energy and zero crossing rate was used.

The model training was done in an office environment, while in the SV testing phase, the audio signal was corrupted by a Gaussian additive noise. The obtained ERR for different SNR in the range of 0 to 30 dB is shown in Figure 5.

According to the obtained results, one can conclude that the conventional SV system is quite sensitive to high Gaussian noise. In noisy environments especially for SNR < 15 dB, noise influence becomes very significant.

4.3. GAD versus VAD (System 2). The quality of the activity detector is measured by the accuracy of speech/nonspeech segment detection. GAD is based on EGG signal and therefore it is robust with the audio noise. Since the EGG signal is only informative during glottal oscillations, GAD detects voiced speech segments. On the other hand, VAD detects both, voiced and unvoiced segments, and uses noise level adaptive threshold causing the narrowing of the detected segments for the increasing noise level.

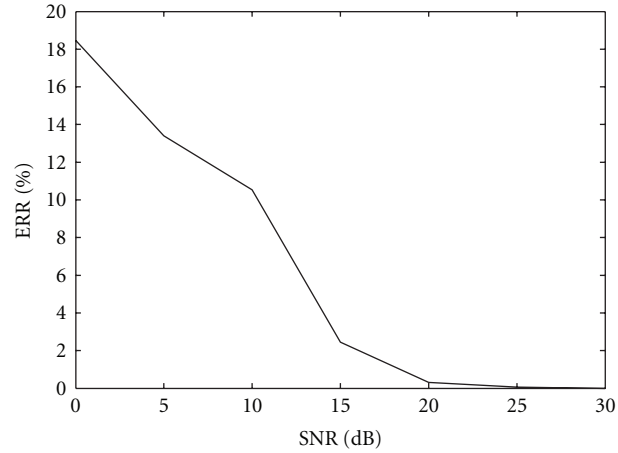


Figure 5: Speaker verification error for different SNR in conventional verification system (System 1).

Figure 6(a) shows a part of natural speech for SNR = 30 dB. Detected segments produced by classical VAD and GAD are denoted by "VAD" and "GAD", respectively. The

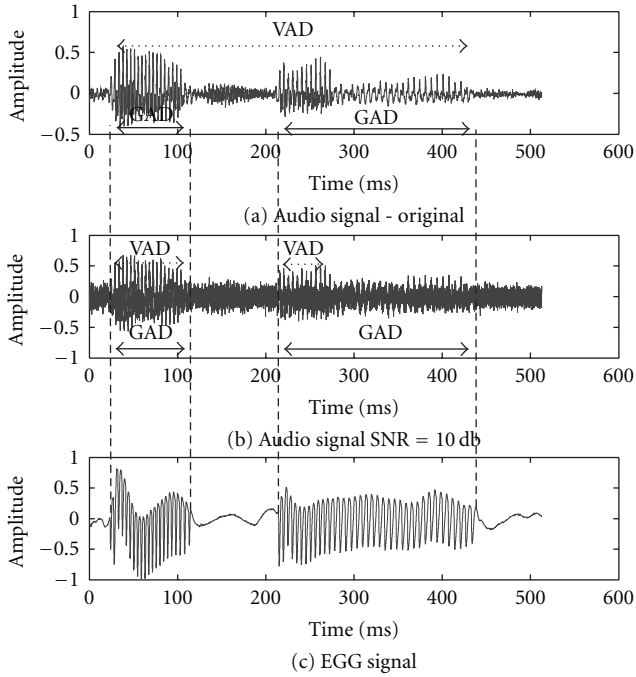


Figure 6: Speech segments obtained by classical VAD and GAD,(a) Audio signal, SNR = 30 dB, (b) audio signal,SNR = 10 dB, (c) EGG signal used for GAD.

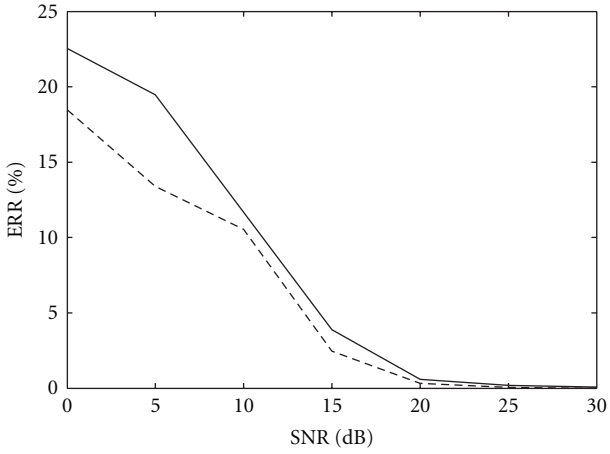


Figure 7: SV error rate using cepstral vectors and GAD (System 2: solid line) and VAD based on energy and zero crossing rate (System 1: dashed line).

same part of the speech for SNR = 10 dB is shown in Figure 6(b). Obviously, detected segments by classical VAD are shorter in Figure 6(b) then in Figure 6(a). At the same time, the effective signal-to-noise ratio is higher for VAD than for GAD. Figure 6(c) shows appropriate EGG signal which is unchanged regardless of SNR value.

The verification system (System 2) used in this experiment was identical as in the previously described System 1, except that VAD was replaced by GAD. The obtained results are plotted as a solid line curve in Figure 7.

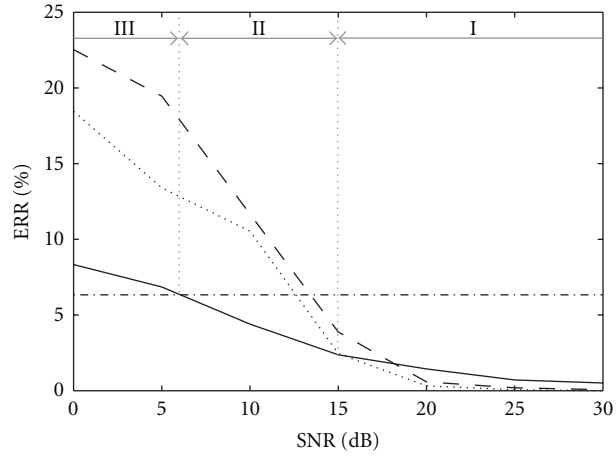


Figure 8: SV error rate using: GAD, EGG features (System 4: dash dot line), GAD, EGG plus audio features (System 3: solid line), GAD, audio features (System 2: dashed line), VAD and audio features (System 1: dotted line).

Although the noise does not have an influence on GAD, the speaker verification error is increased, even in a very noisy environment. Obviously, the results are affected by the choice of detector. VAD separates speech by using adaptive thresholds depending on the level of background noise. On the other hand, GAD-detected speech segments are independent of this level.

4.4. Fusing the EGG Features with Cepstral Coefficients (System 3). In this experiment, conventional feature vectors, $\mathbf{X}_{i,C}$, was augmented by EGG features, $\mathbf{X}_{i,\mathrm{EGG}}$, (altogether 47 coefficients), as in (1). GAD detector was used.

Evaluation and testing was done as in the conventional SV system. The results are shown in Figure 8 as a solid curve.

4.5. Only EGG Features (System 4). Verification system 4 is based only on the feature vectors defined as in (3). GAD was the natural choice as the activity detector. After the enrolment phase, the created GMM models were tested. Considering that EGG feature vectors are not sensitive to audio noise, the obtained result is shown in Figure 8 as a horizontal line, that is, constant value in respect to SNR.

Verification error rates for the different SNR are shown in Table 2. The result presented in the table show benefits δ_1,δ_2 as the difference between conventional SV system 1 and the improved SV systems 3,4.

Throughout the analysis of the results presented here, one can clearly note that the EGG features have a strong influence on the performance of SV in a noise environment. As indicated in Figure 8 and Table 2, substantial gains in speaker verification in a high noise environment were obtained. Analyzing the SNR performance, there are the three different ranges, I, II, and III where System 1, System 3,and System 4 have the best performance, respectively. Therefore, a composite SV can adaptively select one of the three systems, based on the level of noise, achieving a total error rate that is lower than any single system.

Table 2: SV error rate for (I) conventional system, (II) augmented vectors with GAD, (III) only EGG features with GAD and benefits attained, $\delta_1 = (I) - (II)$, $\delta_2 = (I) - (III)$.

	0 dB	5 dB	10 dB	15 dB
(I) System 1	18.46	13.40	10.54	2.45
(II) System 3	8.33	6.85	4.4	2.37
(III) System 4	6.32	6.32	6.32	6.32
δ_1 = (I) − (II)	10.13	6.55	**6.14**	**0.08**
δ_2 = (I) − (III)	**12.14**	**7.08**	4.22	−3.87

One can suggest the use of other speech sensors to create stronger modality combinations that can further be fused using the proposed method to boost the overall performance of an SV system.

These results illustrate the potential of this method for noise robust speaker verification.

5. Conclusion

Considering the sensitivity of noise to a conventional speaker verification system, we examined the informativeness of EGG features. In contrast to the conventional approach, which only extracts cepstral features from audio signal, the proposed method employs information contained within the EGG signal.

The features of the EGG signal, which are robust in a noise environment, are used to augment conventional audio feature vector.

Since EGG signal is only informative during voiced speech segments, the voice activity detector is replaced by a glottal activity detector.

The presented experimental results show a significant reduction of verification error within a noise environment, especially for SNR $<$ 15 dB. As mentioned, there is further improvement, by combining all the systems depending on noise level. Another interesting aspect of the proposed framework is that it could be applied to some other speech modalities by appropriate selection of the activity detector.

As a part of further work, the feature set could be augmented by some other modality which may be more robust against noise, although such a claim would have to be validated. Future work should also explore methods on statistical significant of wider speaker populations to further validate the results.

References

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[2] D. Chow and W. H. Abdulla, "Robust speaker identification based on perceptual log area ratio and gaussian mixture models," in *Proceedings of 8th International Conference on Spoken Language Processing (INTERSPEECH-ICSLP '04)*, pp. 1761–1764, Jeju Island, Korea, October 2004.

[3] P. Premakanthan and W. B. Mikhael, "Speaker verification/recognition and the importance of selective feature extraction: review," in *Proceedings of the 44th IEEE Midwest Symposium on Circuits and Systems (MWSCAS '01)*, vol. 1, pp. 57–61, Ohio. USA, August 2001.

[4] L. Burget, P. Matějka, P. Schwarz, O. Glembek, and J. H. Černocký, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1979–1986, 2007.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[6] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[7] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[8] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP '96)*, pp. 929–932, Philadelpia, Pa, USA, October 1996.

[9] S. Suhadi, S. Stan, T Fingscheidt, and C. Beaugeant, "An evaluation of VTS and IMM for speaker verification in noise," in *Proceedings of the 4th European Conference on Speech Communication and Technology (EuroSpeech '03)*, pp. 1669–1672, Geneva, Switzerland, 2003.

[10] M. J. F. Gales and S. Young, "HMM recognition in noise using parallel model combination," in *Proceedings of the European Conference on Speech Communication and Technology (EuroSpeech '93)*, pp. 837–840, Berlin, Germany, 1993.

[11] T. Matsui, T. Kanno, and S. Furui, "Speaker recognition using HMM composition in noisy environments," *Computer Speech and Language*, vol. 10, no. 2, pp. 107–116, 1996.

[12] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '01)*, vol. 1, pp. 457–460, Salt Lake City, Utah, USA, 2001.

[13] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, pp. 835–838, Munich, Germany, April 1997.

[14] C. Cerisara, L. Rigazio, and J.-C. Junqua, "α-Jacobian environmental adaptation," *Speech Communication*, vol. 42, no. 1, pp. 25–41, 2004.

[15] L. Gonzalez-Rodriguez and J. Ortega-Garcia, "Robust speaker recognition through acoustic array processing and spectral normalization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97)*, vol. 2, pp. 1103–1106, Munich, Germany, 1997.

[16] I. McCowan, J. Pelecanos, and S. Scridha, "Robust speaker recognition using microphone arrays," in *A Speaker Odyssey: The Speaker Recognition Workshop*, pp. 101–106, Crete, Greece, 2001.

[17] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[18] A. Kundu, S. Chatterjee, A. S. Murthy, and T. V. Sreenivas, "GMM based bayesian approach to speech enhancement in signal /transform domain," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 4893–4896, April 2008.

[19] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, pp. 121–124, Seattle, Wash, USA, May 1998.

[20] L. Besacier, J. F. Bonastre, and C. Fredouille, "Localization and selection of speaker-specific information with statistical modeling," *Speech Communication*, vol. 31, no. 2, pp. 89–106, 2000.

[21] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, Article ID 4244529, pp. 1711–1723, 2007.

[22] W. M. Campbell, T. F. Quatieri, J. P. Campbell, and C. J. Weinstein, "Multimodal speaker authentication using nonacoustic sensors," in *Proceedings of the International Workshop on Multimodal User Authentication*, pp. 215–222, Santa Barbara, Calif, USA, 2003.

[23] H. E. Çetingül, E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker/speech recognition using lip motion, lip texture and audio," *Signal Processing*, vol. 86, no. 12, pp. 3549–3558, 2006.

[24] T. F. Quatieri, D. P. Messing, K. Brady et al., "Exploiting nonacoustic sensors for speech enhancement," in *Proceedings of the International Workshop on Multimodal User Authentication*, pp. 66–73, Santa Barbara, Calif, USA, 2003.

[25] B. Zhu, T. J. Hazen, and J. R. Glass, "Multimodal speech recognition With ultrasonic sensors," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH '07)*, vol. 4, pp. 2328–2331, Antwerp, Belgium, 2007.

[26] A. Subramanya, Z. Zhang, Z. Liu, J. Droppo, and A. Acero, "A graphical model for multi-sensory speech processing in air-and-bone conductive microphones," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 2361–2364, Lisbon, Portugal, 2005.

[27] S. Furui, "Survey of the State of the Art in Human Language Technology," 1996, http://cslu.cse.ogi.edu/HLTsurvey/ch1node9.html#SECTION17.

[28] D. G. Chlders, *Speech Processing and Synthesis Toolboxes*, John Wiley & Sons, New York, NY, USA, 2000.

[29] M. Rothenberg and J. J. Mahshie, "Monitoring vocal fold abduction through vocal fold contact area," *Journal of Speech and Hearing Research*, vol. 31, no. 3, pp. 338–351, 1988.

[30] R. J. Baken, "Electroglottography," *Journal of Voice*, vol. 6, no. 2, pp. 98–110, 1992.

[31] M. Hahn and C. K. Park, "An improved speech detection algorithm for isolated Korean utterances," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, pp. 525–528, San Francisco, Calif, USA, March 1992.

[32] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, pp. IV317–IV320, Calif, USA, April 2007.

[33] S. T. Jovicic, Z. Kasic, M. Dordevic, and M. Rajkovic, "Serbian emotional speech database: design, processing and evaluation," in *Proceedings of the 11th International Conference Speech and Computer (SPECOM 04)*, St.Petersburg, Russia, 2004.