A METHODOLOGY FOR SPEAKER-DEPENDENT ACOUSTIC FEATURES BASED ON A SIMPLIFIED CORTICAL RESPONSE FOR SPEAKER VERIFICATION

Guillermo Garcia, Thomas Eriksson

Communication Systems Group, Department of Signals and Systems,
Chalmers University of Technology, 412 96 Göteborg Sweden.
phone: + (46) 7721821, fax: + (46) 7721748
email: em1guill@chalmers.se, thomase@chalmers.se

ABSTRACT

Recently, the incorporation of the research done in biologically inspired systems has shown satisfactory results. A promising research area is the understanding of the human auditory systems and its performance under noisy conditions. Moreover, the incorporation of brain functions (cortical response) as an active part of the auditory system seems a viable alternative to increase the robustness of speech and speaker recognition systems. In this study, we propose a simplified model of the mammalian central auditory system for speaker verification systems. This model is based on a dimension-expansion representation, attempting to capture the response of the cortical cells. Then, by means of the Principal Component Analysis (PCA) approach, we reduce the dimensionality and create speaker-dependent features. Our results showed that by using our modeling technique, we were able to improve the performance of speaker verification systems.

Index Terms— speaker recognition, biological system modeling, feature extraction, pattern classification, robustness.

1. INTRODUCTION

Speaker recognition can be defined as the process of automatically recognizing who is speaking based on the information provided by speech signals. The main technique is to find a set of features that best represents a specific speaker voice. Speaker recognition systems can be classified depending on their tasks in speaker identification (SID) and speaker verification (SV) [1]. In this work, we will focus on SV used to validate whether the speaker is who he or she claims to be.

The speaker recognition process can be divided into two phases independently of the task: enrollment and classification. In the enrollment phase, the expectation maximization (EM) algorithm [2] is used to estimate a Gaussian Mixture Model (GMM) for each speaker enrolled in the database. The EM algorithm provides maximum-likelihood (ML) estimates for the unknown model parameter through a training database. In the classification phase, we compute a score based on the likelihood of test speech samples belonging to a certain speaker. Then, based on the score the speaker recognizer will emit a decision if the speaker is accepted or rejected.

In speech and speaker recognition, the corruption of the speech signals by noise is one of the biggest challenges. Several methods address the problem and attempt to compensate for session/channel/time and microphone variability. Among those methods are cepstral mean substraction (CMS) [3, 4], feature normalization [5], statistical estimation of speech features [6] and many other feature transformations. Moreover, some studies have focused on mimicking and modeling the functions of the human auditory system [7].

During the last years, most of the research done in the area of speaker and speech recognition has focused on modeling the peripheral auditory system without considering the processing stages in the central auditory system. The central auditory modeling was first studied in [8], aiming to create a physiological model of the mammalian auditory system. This auditory models consists of two main parts. The first part is the early auditory model that simulates the processing at the periphery and produces an auditory spectrum, and the second which models the auditory cortex in the central auditory system. In this case, each neuron assumes a response area tuned to a specific range of tone frequencies and intensities, producing a dimension-expanded representation defined as cortical response [9]. The model presented in this study is well-defined and is based on the notion of dimension expansion, where the frequency components of the input signal are mapped to a more redundant representation in the central auditory system. The purposes of using this representation is to obtain information that the peripheral auditory systems are unable to extract and the inclusion of the brain functions as an active part of the recognition process.

In this work, we propose a simplified model of the central auditory system, focusing on the creation of speaker-dependent acoustic features. This approach consists on reducing the dimensionality of the expanded set of features based on the statistical characteristics of each speaker enrolled in the database.

The rest of the paper is organized as follows: section 2 presents an overview of the auditory system, section 3 describes the simplified cortical response, section 4 illustrates the speaker verification framework, section 5 describes the experimental setup of our speaker verification system, section 6 and 7 show the results and conclusions of this work.

2. AUDITORY SYSTEM

The auditory spectrum is a spectral representation produced by an early auditory system consisting of transformations done in the peripheral auditory system, from the ear to the cochlear nucleus. The auditory systems receives a speech signal and passes through a bank of cochlear filters. In this work, the transformations done in the peripheral auditory system are modeled as 3 different stages: pre-emphasis, windowing (usually Hamming) and power spectrum of the speech signal usually obtained from the magnitude of the Discrete Fourier Transform (DFT) of the speech signal defined as

$$\hat{S}_k = \mathbf{F}(s_r) = \sum_{r=1}^{R} s_r \exp\left(\frac{-j2\pi rk}{N}\right) \quad 0 \le k \le R-1, \qquad (1)$$

$$|\hat{S}_k| = \sqrt{\left(\mathrm{Re}\left(\hat{S}_k\right)\right)^2 + \left(\mathrm{Im}\left(\hat{S}_k\right)\right)^2}, \qquad (2)$$

where s_r is the input speech signal after being pre-emphasized and windowed, k represents the discrete frequency variable, and $|\hat{S}_k|$ the absolute value of the DFT. After these transformations, the extracted features are processed by the primary auditory cortex area of the brain.

3. SIMPLIFIED CORTICAL RESPONSE

In the primary auditory cortex, the auditory spectrum is encoded by a population of cortical cells, each of which is characterized by a neural response area that represents the amount of excitation induced by different frequencies [8]. The response areas are organized along three dimensions: central frequencies, scale (bandwidth) and phase. The central frequencies denote the frequencies at which the neurons are excited, the scale denotes the spread area of the response and the phase parameterizes the symmetry.
Figure 1 shows a three dimensional conceptual representation of the cortical response. In this work, we will focus only on the modeling

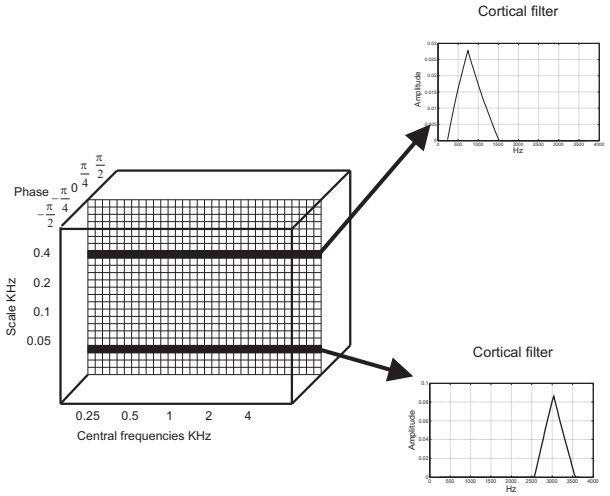

Fig. 1. Conceptual representation of the cortical response.

of the central frequencies and the scale. We assume that the response of each of these filters is triangular and is based on a well-known implementation, and a way of comparison with the Mel Frequency Cepstral Coefficients (MFCC). The simplified cortical response is given by

$$x_n(i) = \sum_{k=1}^{\frac{R-1}{2}} \mathbf{A}_k^{(i)} |\hat{S}_k|, \quad 1 \le i \le M, \qquad (3)$$

where $|\hat{S}_k|$ is the power spectrum defined between $0 \le k \le \dfrac{R-1}{2}$, R is the number of points for the magnitude spectrum, $\mathbf{A}_k^{(i)}$ is the sample magnitude response of the i triangular filter, and M is the total number of filters in the filter-bank. M is defined as $M = M_1 * M_2$ where M_1 is the number of mel central frequency channels and M_2 is the number of scale channels.

Using the previous representation, we obtain an expanded set of features with a larger number of dimensions. The dimensionality of this new set of features yields a computational memory problem. To tackle this problem, we use the Principal Component Analysis (PCA) approach. The PCA is defined as an orthogonal linear transformation of the data to a new coordinate system where the new components are ordered descendingly depending on the greatest variance.

For our specific case, we compute the PCA from the sample covariance matrix of each speaker enrolled in the database. In order to obtain the sample covariance matrix is necessary to calculate first the sample mean defined as

$$\mu_x = \frac{1}{N} \sum_{n=1}^{N} x_n, \qquad (4)$$

where x_n is a feature vector and N is the total number of feature vectors.

Then, the sample covariance matrix is defined as

$$\mathbf{C}_x = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \mu_x)(x_n - \mu_x)^T. \qquad (5)$$

From (5), we can obtain the PCA components by assuming the transformation $y = \mathbf{P}x$ with diagonal covariance matrix $\mathbf{C}_y$ defined as

$$\mathbf{C}_y = \mathbf{P}\mathbf{C}_x\mathbf{P}^T, \qquad (6)$$

where $\mathbf{P}$ are the principal components of x.

Considering j components (row-vectors) of $\mathbf{P}$, we can transform and reduce the dimensionality of the feature set by using the transformation $y = \mathbf{P}x$. Moreover, $\mathbf{P}$ defines a speaker-dependent transformation matrix. To reduce our cortical response feature set, we consider the components with higher variances.

Figure 2 shows the process of the *Simplified Cortical Response* (SCR) from speech samples to the speaker-dependent feature set.

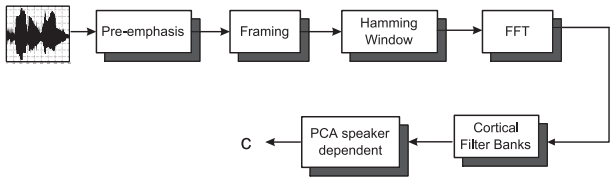

Fig. 2. Diagram of the Simplified Cortical Response.

4. SPEAKER VERIFICATION FRAMEWORK

SV is a statistical hypothesis test between two hypothesis [10]. $H0$ denotes the hypothesis to accept an utterance $\{x_t\}_{t=1}^T$ as being produced by the target speaker. $H1$ denotes the hypothesis to reject an utterance $\{x_t\}_{t=1}^T$ as being produced by the target speaker. Each trial consists of a test utterance and a claimed identity. From each trial a log-likelihood ratio is computed and a score θ is determined as

$$\theta = \ln\left(\frac{p(x|H0)}{p(x|H1)}\right); \qquad \theta \begin{array}{c} \text{accept} \\ \geq \\ < \\ \text{reject} \end{array} \tau. \qquad (7)$$

where τ is the threshold that minimizes the expected cost for errors. The greater the score obtained, the more likely that the trial is the target speaker. $H0$ represents the Gaussian Mixture Model (GMM)

for the target model, and $H1$ represents the impostor model. The impostor model is better known as the universal background model (UBM) which is trained using the information from a pool of speakers different from the target database.

In this work, we consider the EER (equal error rate) and the DET curve[11] as the performance measure.

5. EXPERIMENTAL SETUP

The experiments were conducted using the female speakers from the 2004 NIST-SRE "core" corpus [12]. Each speech file, after removing silence at the beginning and end, was segmented into frames of 25 ms length with an overlap of 10 ms. Each frame was pre-emphasized and Hamming windowed. Then, we implement our SCR over the telephone bandwidth (300-3400 Hz) using 134 mel central frequency channels and 16 scale channels. The scale channels are defined in mel scale as follow

Bw/mel		Bw/mel		Bw/mel		Bw/mel	
1	934.8387	5	545.8816	9	296.1308	13	152.5265
2	823.6421	6	471.3595	10	251.8695	14	128.4962
3	721.7854	7	405.2574	11	213.6092	15	108.0720
4	629.2541	8	347.0471	12	180.7024	16	90.7644

The reason for choosing these frequency bands is that humans can not perceive any difference on a 20% range of the central frequency. After obtaining the SCR, we reduce the dimensionality of the features by applying the PCA for each speaker enrolled in the database. By selecting the 13, 20, and 25 highest components, we create the speaker-dependent set of features for the background model, training and testing.

Afterwards, we train a 512 mixture component speaker-dependent background model, and a speaker model using MAP adaptation [2]. For comparison purposes, we use MFCCs as baseline system. As in the case of the SCR, each speech file, after removing silence at the beginning and end, was segmented into frames of 25 ms length with an overlap of 10 ms. Each frame was pre-emphasized and Hamming windowed. Then 13-th (truncated from 23-th) order MFCCs were created. The training and the evaluation are similar to the cortical response.

6. EXPERIMENTAL RESULTS

Figure 3 shows an example of the cortical filter-bank. As mentioned in the previous section, the SCR considers the frequency and the scale channels. In this figure, we show the filters for three different scale channels, each of them with 135 central frequency channels .

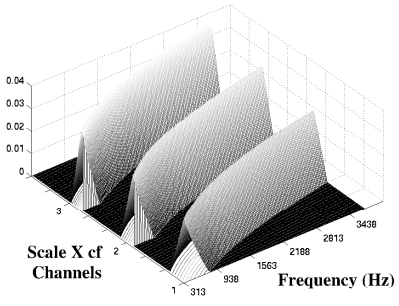

Fig. 3. Example of the cortical filter-banks.

Figure 4 presents an example of the cortical response to a speech frame input. As mentioned in the experimental setup, 134 central frequency and 16 scale channels were used for our experiments. These channels are defined in the horizontal and the vertical axis, respectively.

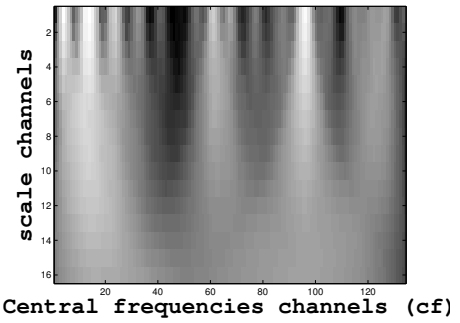

Fig. 4. Response of the neurons to the input of one frame of speech.

Figure 5 shows the DET curves for the MFCCs and the SCR with different components. We observe and improvement as the number of components increases, contrary to the MFCCs case. Table 1

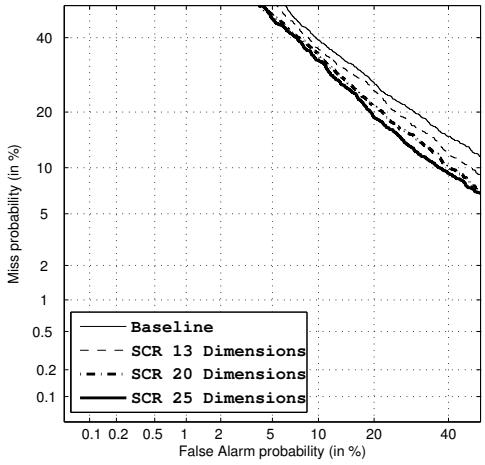

Fig. 5. DET curves for the baseline and the Simplified Cortical Response.

shows the EER for each system. We observe that our SCR outperforms the MFCCs as the number of PCA components increases. A similar increase in dimensions for the MFCCs usually degrades the performance of the system. Although the baseline has a high EER,

Baseline	SCR 13D	SCR 20D	SCR 25D
23.78%	22.51%	20.76%	19.61%

Table 1. EER comparison.

we must consider that the purpose of applying the cortical response is to explore its pure effectiveness as a new set of speaker-dependent features compared to the MFCCs and not to compete with a full-operational SV system. SCR features are different from MFCCs;

hence the methods developed to improve the performance of the MFCCs are not valid for the SCR features.
Figure 6 shows the DET curves for MFCCs (baseline) and the 25 component SCR warped features [5]. We observe that for this case the SCR performance is poor due to the warping process applied. Further study is required to integrate the SCR with all the components of the SV system.

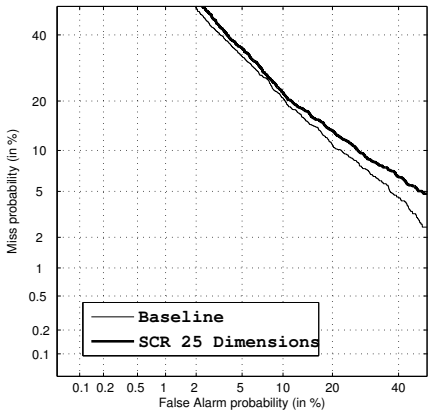

Fig. 6. DET curves for the baseline and the SCR (Warped features).

7. CONCLUSIONS

In this work, we developed a set of speaker-dependent features based on a biologically inspired system. This new set of features considers the neurons response in order to achieve more robustness against the effects of the noise. Using our set of features, we were able to improve the performance of our system. This is an ongoing research and requires further study in the sense of incorporating all the additional elements proper for a baseline SV system and other dimensionality reduction methods.

8. REFERENCES

[1] Douglas Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, 2000.

[2] D. Reynolds and R.C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.

[3] Hynek Hermansky and Nelson Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[4] Steven F. Boll, "Suppresion of acoustic noise in speech using spectral substraction," *IEEE Transactions Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[5] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," in *Proceeding Odyssey*, 2001, pp. 213–218.

[6] Y. Ephraim and M. Rahim, "On second-order statistics and linear estimation of cepstral coefficients," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 2, pp. 162–176, 1999.

[7] M. Holmberg, D. Gelbart, and W. Hemmert, "Automatic speech recognition with an adaptation model motivated by auditory processing," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 43–49, 2006.

[8] X. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Transactions on Speech and Audio Process.*, vol. 3, no. 5, pp. 382–395, 1992.

[9] Woolay Jeon and Biing-Hwang Juang, "Speech analysis in a model of the central auditory system," *IEEE Transactions on Speech and Audio Process.*, vol. 15, no. 6, pp. 1802–1817, 2007.

[10] Frederic Bimbot, Jean-Francois Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega Garcia, Dijana Pretrovska, and Douglas Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, pp. 430–451, 2004.

[11] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki, "The det curve in assessment of detection task performance," in *Proceeding EUROSPEECH*, 1997, pp. 1985–1898.

[12] "The NIST 2004 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/2004/.