A LOGARITHMIC BASED POLE-ZERO VOCAL TRACT MODEL ESTIMATION FOR SPEAKER VERIFICATION

Ewald Enzinger[1], *Peter Balazs*[1], *Damián Marelli*[2], *Timo Becker*[3]

[1] Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria
[2] School of Elect. Engineering and Computer Science, University of Newcastle, Australia
[3] Federal Criminal Police Office, Germany

ABSTRACT

In this paper we investigate the use of formant and anti-formant measurements of nasal consonants for speaker verification. The features are obtained using a pole-zero vocal tract model estimate optimized by minimizing a logarithmic criterion which is motivated by the perception of amplitude by the human auditory system. A GMM-UBM approach is used for performing speaker comparisons within the likelihood-ratio framework. Results are compared with systems based on Mel Frequency Cepstral Coefficients (MFCCs) as well as formant center frequencies and bandwidths obtained using the Snack Toolkit. The formant and anti-formant based system attains comparable results to the MFCC system and outperforms the formant-based approach while offering a more straightforward interpretation in terms of a physical speech production model.

Index Terms— Speaker recognition, speech analysis, pole-zero model, formants, anti-formants

1. INTRODUCTION

Automatic speaker verification systems, especially those targeted at the forensic field, predominantly use the Gaussian Mixture Model – Universal Background Model (GMM-UBM) approach, combined with cepstral features such as Mel Frequency Cepstral Coefficients (MFCC). While this combination of classifier and features provides good performance, the lack of a straightforward interpretation of these features with regard to a physical model of vocal tract properties of a speaker leaves them as an unfavorable choice for certain applications such as providing evidence to the court.

On the other hand, formant features as they are used in acoustic-phonetic approaches to forensic speaker comparison [1] can be related to the resonance cavities of the vocal tract. Formant center frequencies and their bandwidths are sufficient to determine the areas of an acoustic tube formed by cascading M uniform cylindrical sections of equal length [2]. They have been successfully applied to the task of forensic speaker comparison using the GMM-UBM approach [3].

Formants are usually measured by methods based on all-pole models of the speech production filter which provide a good characterization of some speech sound categories. Representations of the vocal tract for unvoiced and nasal as well as lateral sounds contain the anti-resonances (zeros) and resonances (poles) of the vocal tract. Therefore, pole-zero models offer an advantage. Here we present the estimation method described in [4].

2. POLE-ZERO MODEL

Speech production is modeled by a linear, slowly time-varying filter, the speech production filter (SPF), which models the combined effect of the vocal tract and the radiation of the lips, as well as the glottal pulse shape in the case of voiced sounds. It is assumed to be time-invariant during a short-time period of approximately 20-40ms.

In the speech production model, the sampled speech signal $y(t)$ is assumed to be generated by an excitation signal $u(t)$ filtered by the SPF $g_t(\tau)$, i.e.

$$y(t) = \sum_{\tau \in \mathbb{Z}} g_t(\tau) u(t-\tau). \tag{1}$$

The signal $u(t)$ is assumed to be a train of impulses for voiced sounds, or white noise in the case of unvoiced sounds.

The frequency response of the SPF is given by

$$G(z,\theta) = \frac{B(z,\theta)}{A(z,\theta)} = \frac{\sum_{l=0}^{n} b_l z^{-l}}{\sum_{l=0}^{m} a_l z^{-l}}, \tag{2}$$

where n and m denote the orders of the numerator and denominator, respectively. The set of parameters θ denoted as

$$\theta = [b_0, b_1, \cdots, b_n, a_1, \cdots, a_m]^T \tag{3}$$

is tuned to fit a frequency response estimate $\hat{G}(\omega_k)$ of the SPF at a discrete set of frequencies $\{\omega_k, k = 1, \cdots, K\}$. The present work uses the method described in [5] which obtains $\hat{G}(\omega_k)$ by interpolating spectral peaks found within neighborhoods of the multiples of the pitch frequency.

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

Motivated by the fact that the human auditory system is perceiving amplitude of the frequency contents of a sound signal in a logarithmic scale [6], the coefficients are optimized by minimizing the following logarithmic criterion:

$$\theta = \arg\min_{\theta'} \sum_{k=1}^{K} \left| \log |\hat{G}(\omega_k)| - \log \left| \frac{B(e^{j\omega_k}, \theta')}{A(e^{j\omega_k}, \theta')} \right| \right|^2 . \quad (4)$$

The optimization problem (4) can be written as

$$\theta = \arg\min_{\theta'} V(\theta'), \quad (5)$$

$$V(\theta) = \sum_{k=1}^{K} [F(\theta)]_k^2, \quad (6)$$

where $[F(\theta)]_k$ denotes the k-th component of the real-valued vector $F(\theta)$, which is a function of the d-dimensional real-valued vector θ. Then, (5)-(6) are equivalent to (4) if we define

$$[F(\theta)]_k = \log \left| \frac{\hat{G}(\omega_k)}{G(e^{j\omega_k}, \theta)} \right|, \text{ for all } k = 1, \cdots, K, \quad (7)$$

Using Newton-like methods, (5)-(6) is solved using the following iterative procedure

$$\theta_{i+1} = \theta_i - \alpha_i \tilde{\theta}_i, \quad (8)$$

where $\tilde{\theta}_i$ is the solution of

$$H_i \tilde{\theta}_i = g_i, \quad (9)$$

the scalar α_i denotes the step size at iteration i, the d-dimensional vector g_i denotes the gradient of $V(\theta)$ at θ_i, and the $d \times d$ matrix H_i denotes either the Hessian of $V(\theta)$ at θ_i or an approximation of it.

Let $J(\theta)$ denote the Jacobian of $F(\theta)$, i.e.,

$$[J(\theta)]_{k,l} = \frac{\partial [F(\theta)]_k}{\partial [\theta]_l}. \quad (10)$$

The gradient g_i can be computed from the Jacobian information by

$$g_i = 2J^T(\theta_i)F(\theta_i). \quad (11)$$

We use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula [7], a iterative procedure that directly approximates H_i^{-1}:

$$\begin{aligned} H_{i+1}^{-1} &= H_i^{-1} + \left(1 + \frac{q_i^T H_i^{-1} q_i}{s_i^T q_i}\right) \frac{s_i s_i^T}{s_i^T q_i} \\ &\quad - \frac{s_i q_i^T H_i^{-1} + H_i^{-1} q_i s_i^T}{s_i^T q_i}, \\ s_i &= \theta_{i+1} - \theta_i, \\ q_i &= g_{i+1} - g_i. \end{aligned}$$

The step-size parameter α_i is obtained from a linear search algorithm using a sub-iterative procedure (i.e., formed of *sub-iterations* of the *main iterations* (8)-(9)) in which, starting from the initial value $\alpha_i = 1$, the value of α_i is halved at each sub-iteration until

$$V(\theta_i - \alpha_i \tilde{\theta}_i) < V(\theta_i),$$

or a maximum number of iterations is reached.

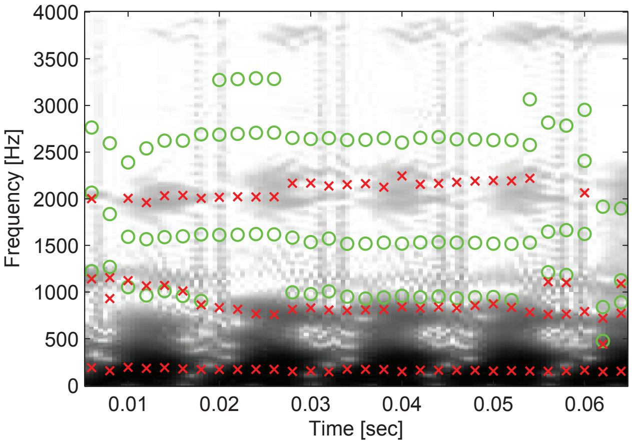

Fig. 1. Formant (x) and anti-formant (o) measurements over the length of a /n/ consonant

The formant and anti-formant measurements are obtained from the roots of the denominator and numerator polynomials in the z-plane, sorted in ascending order. The signal is divided into frames using a 40ms hanning window and 95% overlap. An order of 11 is selected for both numerator and denominator, which was determined based on a subset of the data. The set of coefficients is first initialized by a weighted linear least-squares algorithm [8] and then optimized by the proposed method. The procedure does not employ any tracking algorithm, i.e. any consideration of the temporal inter-relation of the estimated poles and zeros, and imposes no continuity conditions on obtained values. Fig. 1 shows an example of formant and anti-formant measurements of an /n/ sound.

3. SPEAKER VERIFICATION SYSTEM

The automatic speaker verification system used in this study is based on the GMM-UBM approach [9] and extends previous work in [3, 10] where it was applied to formant center frequencies and bandwidths. Feature vectors consist of the first three formants and anti-formants, as determined by the pole-zero model, yielding 6 features per frame. Speakers are modeled by a Gaussian mixture model (GMM) denoted by

$$\lambda := (p_i, \mu_i, \Sigma_i)_{i=1,\ldots,M}, \quad (12)$$

where p_i, μ_i and Σ_i represent the mixture weights, means and covariance matrices. The *universal background model* (UBM) is created by training a GMM from pooled feature vectors of different speakers using a maximum likelihood criterion, which is solved using the Expectation-Maximization (EM) algorithm. Speaker models are derived from the UBM using maximum a-posteriori (MAP) adaption. This was found to provide better results than the original approach in [3, 10]. A number of 8 mixture components is used in accordance to [10]. Full covariance matrices are used in order to be able to properly model within-speaker variability. The likelihood of a set of feature vectors X given a GMM λ is calculated by

$$P(X|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda). \tag{13}$$

where $f(x_i|\lambda)$ is the Gaussian mixture density function for the specified model λ. In each speaker comparison the likelihood ratio (LR) is computed for a set of test feature vectors X and the models of a speaker and the UBM.

$$LR = \frac{P(X|\lambda_{speaker})}{P(X|\lambda_{UBM})} \tag{14}$$

This score usually does not represent a proper LR which requires that same-speaker comparisons report high LR values while different-speaker comparisons report low values, with values close to one offering no support to any of the two hypotheses. Therefore, an automatic calibration procedure based on logistic regression is applied to the scores using the methods provided by the FoCaL toolkit [11]. The parameters are estimated in a cross validation setting, using the scores of all speakers except those involved in the current trial.

4. EVALUATION

Performance comparisons are carried out using the proposed method, a baseline GMM-UBM speaker verification system using basic MFCC features which is described in Section 4.1 as well as an approach using formant center frequencies and bandwidths which are extracted using the Snack Toolkit[1] (subsequently denoted as *Formants/BW*). This approach is akin to [3], but uses MAP adaption to obtain speaker models.

All three systems are applied to the same /n/ consonant data which is described in Section 4.2. The configuration of both systems and features are chosen as example for this data in line with previous work on speaker verification [3, 10, 9]. Their optimization will be dealt with in future work.

The equal error rate (EER) and the *log likelihood-ratio cost* (C_{llr}) metric [11] are used as performance measures. Detection error trade-off (DET) plots are used to show the trade-off between type I and II errors when the decision threshold is varied over the LR range. Tippett plots characterize the cumulative proportion of LRs from target trials less than or equal to the value indicated on the abscissa and of non-target trials greater than or equal to the value on the abscissa.

4.1. Baseline system description

A GMM-UBM system using Mel Frequency Cepstral Coefficients (MFCCs) [9] is used as baseline to compare speaker verification performance. Feature vectors of 13 MFCCs are computed every 10ms using a 20ms hamming window. After extraction, cepstral mean reduction (CMR) is applied to the feature vectors. The system is based on Gaussian mixture models with 1024 mixture components and diagonal covariance matrices[2]. Models of individual speakers are obtained through MAP adaption from the UBM. No further score normalizations such as the T-norm are applied.

4.2. Data base

The evaluations in this study are based on nasal /n/ consonants in recordings of 106 male adult German speakers which were selected from the Pool2010 corpus [12]. To obtain a sufficient number of items, an automatic phone-level alignment [13] was performed on recordings of the German version of *the north wind and the sun* read by the speakers in one studio session. Subsequently, auditory validation of the segments was performed to check for possible alignment errors.

30 speakers were used for UBM training. This number was chosen based on the results in [10]. The data of the remaining 76 speakers was split into two equally-sized train and test datasets of about 25 /n/ segments with a median duration of 60ms, allowing for 76 target and 5700 non-target trials.

5. RESULTS AND DISCUSSION

Table 1 provides the EER and C_{llr} values of the different systems. The proposed method provides discrimination performance in terms of EER equal to the MFCC based system and outperforms the formant features. In terms of C_{llr}, it incurs a slightly higher cost than the baseline system and a lower cost than the formant-based systems. In the DET plot in Fig. 2 the proposed method displays similar characteristics as the MFCC system except for thresholds minimizing the false alarm rate. The Tippett plot which is of interest in the context of forensics is shown in Fig. 3.

Features	EER	C_{llr}
proposed method	3.9%	0.1325
Formants/BW	5.3%	0.2226
MFCC	3.9%	0.1296

Table 1. EER and C_{llr} results

6. CONCLUSIONS

In this paper, a new set of features consisting of formant and anti-formant measurements obtained from a logarithmic based pole-zero model estimate of the speech production filter [4] is applied to the task of speaker verification. These

[1] http://www.speech.kth.se/snack/

[2] A similar configuration was used in [9] for single-gender UBMs.

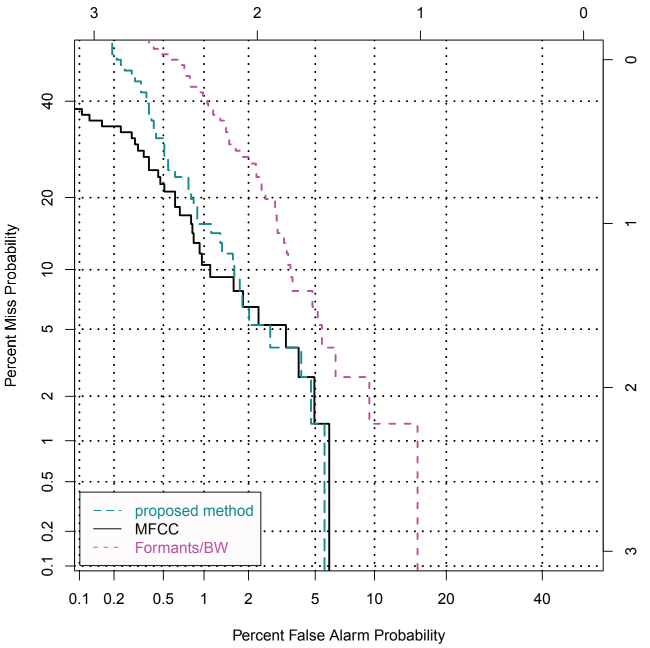

Fig. 2. DET plot of the compared systems

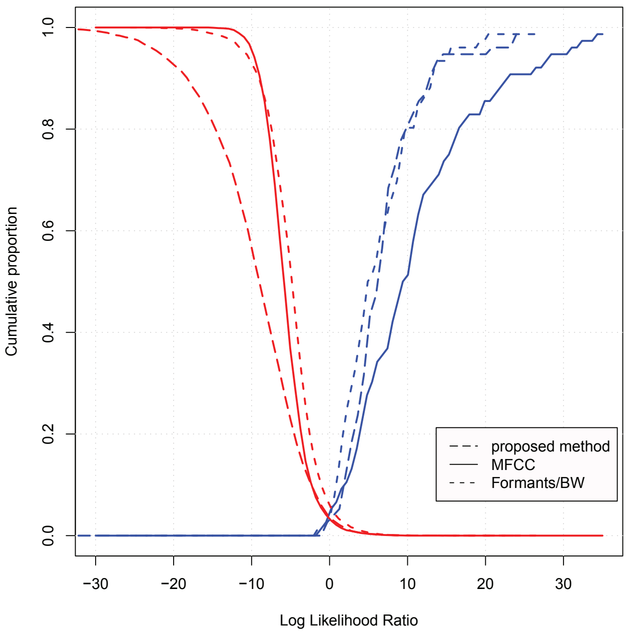

Fig. 3. Tippett plot of the compared systems

features are advantageous due to their more straightforward interpretation. The features were extracted in an unsupervised procedure and subsequently used in a GMM-UBM speaker comparison approach. In an evaluation based on nasal /n/ consonants, this set of features achieves performance values comparable to a MFCC based approach and outperforms an approach based on formant frequencies and bandwidths.

Further tests are needed to evaluate the method on non-contemporaneous speech as well as its susceptibility to channel mismatch such as transmission over telephone using speech codecs and differences in speaking style and duration, which is especially important for forensic applications [14].

A further improvements of the proposed method could be achieved by using perceptual frequency scale as it is applied in Perceptual Linear Prediction (PLP) and by adding derivatives of the features, as commonly performed on MFCCs in speaker verification systems. Furthermore, the amount of complementary information to MFCC/PLP features and the order of improvement achievable through fusion techniques needs to be investigated.

7. REFERENCES

[1] P. Rose, *Forensic Speaker Identification*, Taylor & Francis, 2002.

[2] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” *J. Acoust. Soc. Amer.*, vol. 50, no. 2B, pp. 637–655, 1971.

[3] T. Becker, M. Jessen, and C. Grigoras, “Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models,” in *Proc. Interspeech*, Brisbane, 2008, pp. 1505–1508.

[4] D. Marelli and P. Balazs, “On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 2, pp. 237–248, Feb. 2010.

[5] H. Hermansky, H. Fujisaki, and Y. Sato, “Spectral envelope sampling and interpolation in linear predictive analysis of speech,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 1984, vol. 9, pp. 53–56.

[6] W. Hartmann, *Signals, Sounds and Sensation*, Springer, New York, 1998.

[7] R. Fletcher, *Practical Methods of Optimization*, ser. A Wiley-Interscience Publication. Wiley, Chichester, U.K., 2nd ed., 1987.

[8] T. Kobayashi and S. Imai, “Design of IIR digital filters with arbitrary log magnitude function by WLS techniques,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 2, pp. 247–252, Feb. 1990.

[9] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[10] T. Becker, M. Jessen, and C. Grigoras, “Speaker verification based on formants using gaussian mixture models,” in *Proc. NAG/DAGA*, Rotterdam, 2009.

[11] N. Brümmer and J. du Preez, “Application-independent evaluation of speaker detection,” *Comput. Speech Lang.*, vol. 20, pp. 230–275, 2006.

[12] M. Jessen, O. Köster, and S. Gfroerer, “Influence of vocal effort on average and variability of fundamental frequency,” *Int. J. Speech, Language, and the Law*, vol. 12, no. 2, pp. 174–213, 2005.

[13] S. Rapp, “Automatic phonemic transcription and linguistic annotation from known text with hidden markov models,” in *Proc. ELSNET Goes East and IMACS Workshop*, Moscow, 1995.

[14] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. F. Bonastre, and D. Matrouf, “Forensic speaker recognition: A need for caution,” *IEEE Signal Processing Magazine, Special Issue on Digital Forensics*, vol. 26, no. 2, pp. 95–103, 2009.