Experiments on using Vocal Tract Estimates of Nasal Stops for Speaker Verification

Ewald Enzinger, Christian H. Kasess
Acoustics Research Institute
Austrian Academy of Sciences
Vienna, Austria
{ewald.enzinger, christian.kasess}@oeaw.ac.at

***Abstract*— Nasal stops have been recognized as an important source of speaker-discriminating features. The nasal cavity is, with the exception of the velar junction, independent of articulatory movements. As the complex nasal structure varies from person to person, features dependent upon nasal acoustics may have low within-speaker and high between-speaker variability. In this study we use a Bayesian estimation technique to obtain reflection coefficients of a branched-tube model of the combined nasal and oral tract. These are then used as parameters in speaker verification experiments. The performance is evaluated on the basis of speakers from the TIMIT corpus as well as the Kiel corpus and is compared with that of a system based on Mel frequency cepstral coefficient (MFCC) features. Fusion of both systems indicates that the two approaches offer complementary information.**

***Keywords*— Nasals, vocal tract modeling, Bayesian estimation, speaker verification**

I. Introduction

Nasals have been considered to potentially provide useful information for discriminating speakers [1] (p. 133). In the production of nasal stops the nasal cavity is coupled to the vocal tract by lowering the velum while a closure is formed by the lips (/m/), the tongue at the alveolar ridge (/n/) or the tongue dorsum at the lowered velum (/ŋ/). The relatively fixed structure of the vocal and nasal cavity provides the basis for the a-priori assumption of low within-speaker variability.

During closure, the pharynx and the nasal cavities form a pathway, which acts as a filter for the glottal pulse stream. It causes peaks in the spectrum corresponding to its resonances, while the closed oral cavity introduces peaks as well as depressions that are caused by acoustical cancellations. Pairs of sinuses, the sphenoidal sinus, maxillary sinus, frontal sinus and the ethomoidal sinus, commonly called paranasal cavities, are located around the nasal cavity and are coupled to it, which causes additional resonances and cancellations [2,3]. Due to their complicated structure and the asymmetric proportions of the left and right sinuses and passages of the nasal tract, which is split in two by the nasal septum, there exists substantial variation in the acoustic properties between different speakers [4]. Combined with the low within-speaker variability, the acoustics of nasal stops are theoretically a valuable source of speaker-discriminating features.

The use of nasal segments was demonstrated in early studies on speaker identification [5,6]. In automatic speaker recognition, work on the relative value of different sound classes and representations identified nasal stops as a particularly important source of speaker-discriminating features [7,8,9,10]. Most studies, however, did not explore explicit modeling of nasal acoustics beyond modeling spectra using pole-zero model estimates [11,12]. Features derived from theoretical models of the vocal tract acoustics can more readily be interpreted, which may be beneficial for applications such as forensic voice comparison.

The drawback of such models is the higher complexity and thus a more difficult estimation. To accurately model the spectral components of nasal speech signals, a minimum of two connected tubes is necessary. This added complexity as compared to one-tube models requires additional assumptions in order to constrain the estimation process. Thus, the present paper uses a variational Bayesian scheme to estimate the tube areas of a branched nasal and oral tract model from the log-spectrum of the speech signal of nasal stops [13]. Probabilistic priors are used to enforce smoothness of the tube model. Reflection coefficients are obtained from tube model estimates and are used as features. We use a Gaussian mixture model – universal background model (GMM-UBM) system based on these features in speaker verification experiments on the TIMIT data base and the German-language Kiel corpus. The effect of using different prior variances in the estimation is evaluated. Performance is compared with a baseline GMM-UBM system using Mel frequency cepstral coefficients (MFCCs), a standard feature in automatic speaker recognition systems, which are extracted from the same nasal stop segments.

II. Methodology

A. Data base

We performed speaker verification experiments using two data bases in different languages: The TIMIT corpus [14] and 55 German speakers taken from the Kiel Corpus [15]. These corpora were selected because they provide accurate phonetic labels, which otherwise would have to be acquired by using a nasal detector or a speech recognizer followed by forced alignment. For experiments on the Kiel corpus we selected 20 speakers to train the universal background model. Data from the remaining 35 speakers were used to form verification trials for evaluating the performance of the features. We pooled alveolar nasal /n/ segments from all sentences spoken by a speaker and split them into two sets of approximately equal size for use in model training and as test data in verification trials. For TIMIT we selected each 20 female and 20 male speakers to train the universal background model. Further 20 female and 20 male speakers were selected as development set to optimize the number of mixture components in the Gaussian mixture model and to train calibration weights (see Section II-E). Data from the remaining 550 speakers were used to form verification trials for evaluating the performance of the features. We pooled alveolar nasal /n/ segments from *sa* and *si* sentences for adapting speaker models and alveolar nasal /n/ segments from *sx* sentences as test data.

B. Vocal tract model

Nasal stops can be modeled by an acoustic tube consisting of three parts [16]: (1) a pharyngeal tract (L segments) between the glottis and the velum (nasal-oral branching point), (2) a nasal tract (M segments) open at the nostrils, and (3) a closed (non-radiating) oral tract (N segments). Each tract is modeled by a segmented tube. Using continuity conditions between the segments and at the coupling of the three branches a pole-zero representation or rational transfer function $H(z) = B(z) / A(z)$ can be derived. These polynomials are related to the area function of the vocal tract via the reflection coefficients

$$\mu_m = \frac{A_{m+1} - A_m}{A_{m+1} + A_m}, \tag{1}$$

where A_m is the cross-sectional area of the m-th segment starting at the nostrils (or lips for the oral part). The numerator polynomial $B(z)$ is of degree N and dependent on the oral reflection coefficients $\mu_{0;O}$ to $\mu_{N–1;O}$. The denominator polynomial $A(z)$ of degree $L + M + N$ is dependent on the oral reflection coefficients, the pharyngeal reflection coefficients μ_M to $\mu_{M+\Lambda}$, the nasal reflection coefficients μ_1 to $\mu_{M–1}$, and the relation between the cross sectional areas of oral and nasal coupling sections at the velum [16]:

$$\nu = A_{N-1,O} / (A_{M-1} + A_{N-1,O}) \tag{2}$$

However, in general, no exact mapping from the $M + L + 2N$ polynomial coefficients to the $M + L + N + 1$ tube model parameters exists. Hence, deriving the area function of a branched-tube model from a pole-zero model is not straightforward and requires some degree of approximation [16,17].

In [13] we introduced an approach that estimates all coefficients directly from the spectral envelope. A Bayesian algorithm is used that includes probabilistic prior assumptions on the smoothness of the vocal tract tube. The estimation scheme is based on a general variational Bayesian scheme under Gaussian assumptions [18] and has been shown to decrease the within-speaker variability [13].

C. Estimation scheme

Based on the results of [19], the estimation scheme models the logarithm of the transfer function $H(z)$ based on the log of the spectral envelope $G(\omega)$ of the recorded signal. The generative model for the log-envelope can be written as

$$y = \ln G(\omega) = f(\theta, \omega) + \varepsilon(\omega) \tag{3}$$

The function $f(\theta, \omega)$ incorporates the non-linear transformation from the reflection coefficients to the log transfer function as well as a non-linear mapping from the i-th parameter θ_i (which is unrestricted) to the i-th reflection coefficient μ_i using a sigmoidal function (specifically the Gaussian error function) ensuring that the reflection coefficients are restricted to the open interval (–1, 1). The nasal-oral coupling parameter ν is restricted to the interval (0, 1). A scaling factor for the transfer function is also added. This parameter has to be positive, which is achieved by a log transformation. Therefore, the parameter vector θ is of dimension $M + N + L + 2$. The measurement error ε is assumed to be normally distributed with $N(0, \Sigma(\lambda))$, with λ parameterizing the error covariance such that $\Sigma(\lambda)$ is a diagonal matrix with $\exp(-\lambda)$ as its entries. The normality assumption about the error yields a Gaussian likelihood function

$$p(y \mid \theta, \lambda, m) = N(y \mid f(\theta), \Sigma), \tag{4}$$

where $\Sigma(\lambda)$ is now written as Σ for simplicity and m denotes the model assumptions, e.g., prior settings and vocal tract structure. The priors for θ and λ are also Gaussian, i.e.,

$$p(\theta) = N(\theta \mid \eta_\theta, \Pi_\theta^{-1}) \text{ and } p(\lambda) = N(\lambda \mid \eta_\lambda, \Pi_\lambda^{-1}), \tag{5}$$

where m was dropped for simplicity. Π_θ and Π_λ are the respective precision matrices.

D. Vocal tract priors

Informative priors for the reflection coefficients would require probabilistic information about the vocal tract shape, which are not well known in general. Therefore, we just require smoothness of the vocal tract (a similar example is obtaining the area function using linear predictive coding when both glottal and lip losses are estimated [20]). Solutions with smaller reflection coefficients and hence smoother vocal tracts are preferred by using Gaussian priors centered on zero. A higher prior precision (i.e., a smaller prior variance) implies stronger regularization. The prior for the nasal-oral coupling coefficient ν is also centered on zero resulting in equal nasal and oral coupling areas due to the non-linear sigmoidal mapping.

E. Speaker verification experiments

The Gaussian mixture model – universal background model approach [21] was adopted in the experiments for modeling the extracted features. We favored this technique over current state-of-the-art speaker verification approaches such as i-vector, joint factor analysis, or support vector machine based systems, as these systems require additional training data which is often not easily available, e.g., in forensic applications.

Feature vectors consisting of the transformed reflection coefficients θ_i of vocal tract model estimates were modeled by mixtures of Gaussian distributions (GMMs) with diagonal covariance matrices, denoted by

$$\lambda := (p_j, \tau_j, \Sigma_j)_{j=1,\ldots,K}\,, \tag{6}$$

where p_j, τ_j, and Σ_j represent the mixture weights, means and covariance matrices. The universal background model (UBM), which models the distribution of the features in the reference population, was trained on the background data pooled across speakers. Its mixture weights, means and co-variances were estimated using the expectation-maximization (EM) algorithm. Individual GMMs that represent speakers were obtained by maximum a-posteriori (MAP) adaption of the UBM means. In a comparison trial a score was calculated as

$$s = \frac{1}{N}\sum_{k=1}^{N}\log\left(\frac{p(x_k \mid \lambda_{speaker})}{p(x_k \mid \lambda_{UBM})}\right), \tag{7}$$

where x_k is a feature vector, N is the number of tokens of /n/ in the test data, and $\lambda_{speaker}$ and λ_{UBM} represent the models of the trial speaker and the background model, respectively.

For each test we calculate scores using a GMM-UBM system based on vocal tract reflection coefficients (VT RCs) as feature vectors and a baseline GMM-UBM system based on Mel frequency cepstral coefficients (MFCCs). Scores of each system were calibrated using logistic-regression calibration [22]–[24]. For experiments on the TIMIT corpus a small development set of speakers was used to obtain the calibration weights, for experiments on the Kiel corpus we obtained calibration weights for each comparison score using leave-one-out cross validation (calculations were performed using [25], and [26]). We also examined whether combining the two different types of representation leads to improved performance. For this we used logistic regression to fuse the scores from VT RC and MFCC based systems [27], which is a commonly used fusion technique.

III. RESULTS

A. TIMIT corpus

We first investigated the effect of using different values for the prior precision in the Bayesian estimation scheme on speaker verification performance. Fig. 1 shows a Detection Error Trade-off (DET) plot [28] (obtained using the Receiver Operator Characteristic Convex Hull method [29]) of systems using vocal tract parameters obtained using different prior variances from alveolar nasal /n/ tokens from speakers in the TIMIT corpus. Steady increases in performance can be observed for higher values of the prior precision.

Fig. 2 shows a comparison of the performance of a system based on vocal tract reflection coefficients (VT RCs) and MFCCs extracted from alveolar nasal /n/ tokens from speakers in the TIMIT corpus. The system based on vocal tract reflection coefficients had an EER of 12% and the system based on MFCC an EER of 10.7%. Fusion of both systems achieved a sizable absolute reduction in EER to 7.5%, indicating that both systems offer complementary information.

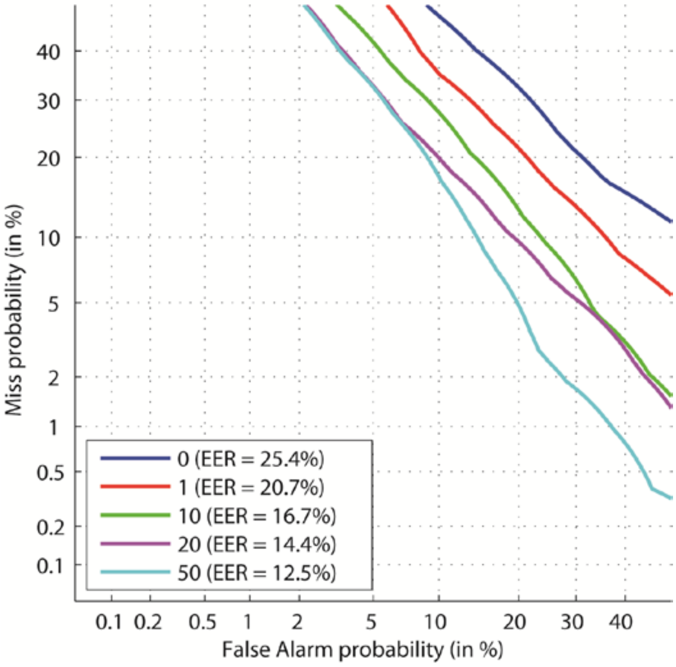

Fig. 1. DET plot comparing the performance of using different prior variances in the Bayesian estimation of vocal tract parameters extracted from alveolar nasal /n/ tokens from speakers in the TIMIT corpus.

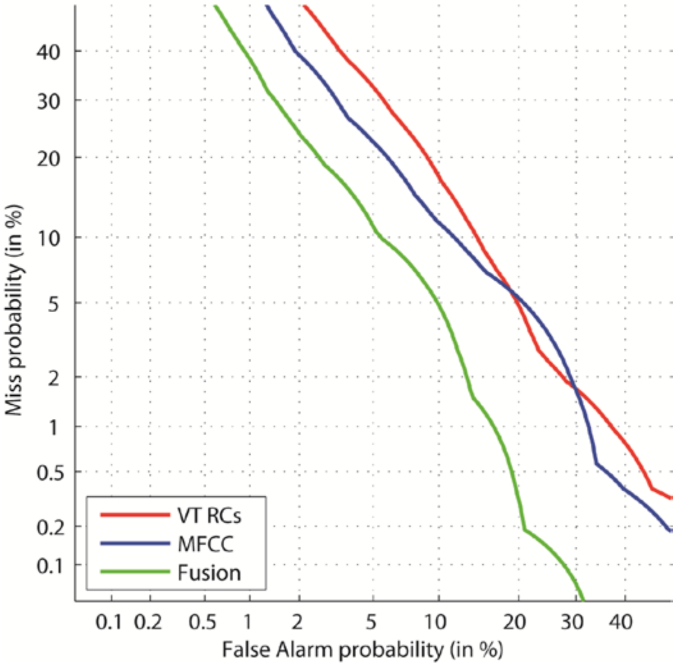

Fig. 2. DET plot comparing the performance of vocal tract reflection coefficients with MFCCs, both extracted from alveolar nasal /n/ tokens from speakers in the TIMIT corpus, as well as fusion of both systems.

B. Kiel corpus

We again first compare performance when using different prior variances in the Bayesian estimation scheme. Fig. 3 shows a ROCCH DET plot of systems using vocal tract parameters obtained by applying different prior variances extracted from alveolar nasal /n/ tokens from speakers in the Kiel corpus.

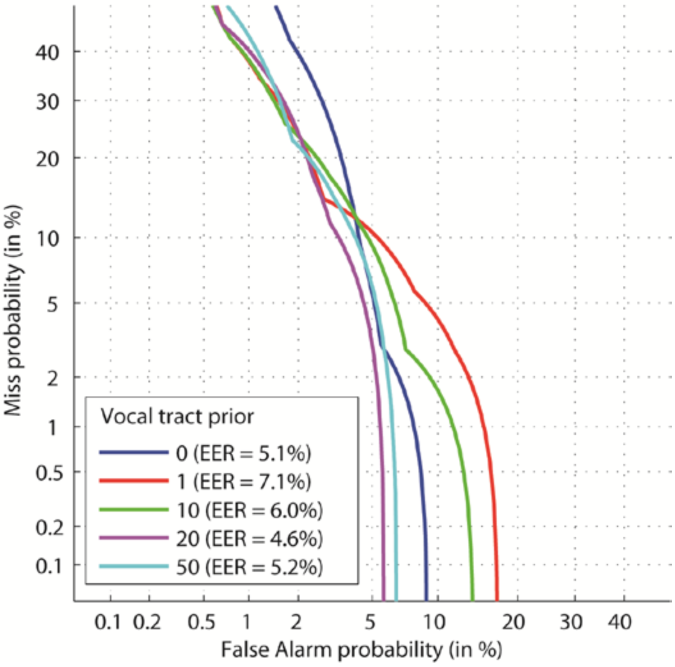

Fig. 3. DET plot comparing the performance of using different prior variances in the Bayesian estimation of vocal tract parameters extracted from alveolar nasal /n/ tokens from speakers in the Kiel corpus.

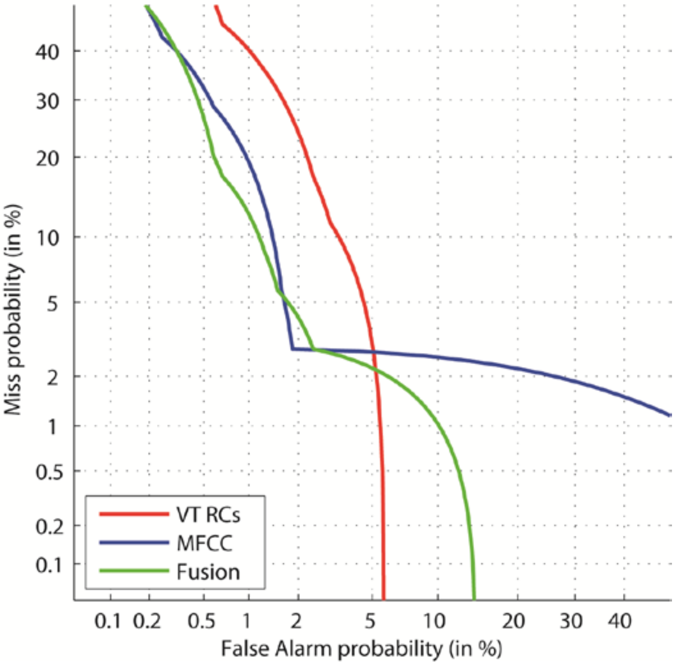

Fig. 4. DET plot comparing the performance of vocal tract reflection coefficients with MFCCs, both extracted from alveolar nasal /n/ tokens from speakers in the Kiel corpus, as well as fusion of both systems.

Here the results were less clear than in the previous experiment. The system using vocal tract parameters estimated using a prior precision of 20 showed the lowest EER at 4.6%.

Fig. 4 shows a comparison of the performance of a system based on vocal tract parameters and MFCCs extracted from alveolar nasal /n/ tokens from speakers in the Kiel corpus. The system based on vocal tract parameters has an EER of 3.4% and the system based on MFCC an EER of 2.8%. Fusion of both systems does not yield a substantial improvement, with an EER of 2.78%. However, due to the low number of speakers and thus target trials in the evaluation the results should be considered with caution, in particular in the low miss probability range.

IV. Discussion and Conclusion

The present paper reports on experiments using physiologically motivated vocal tract parameters modeling both, oral and nasal acoustics as features for speaker verification. A Bayesian estimation technique is used to obtain reflection coefficients of a branched two-tube model of the combined nasal and oral tract during alveolar nasal /n/ segments. Reflection coefficients of the tube model are used as features.

Higher values for the prior precision in the Bayesian vocal tract estimation generally show better performance than lower values (50 for experiments on TIMIT and 20 for experiments on the Kiel corpus). In the experiments reported here the vocal tract parameters have not been explicitly optimized to attain high speaker discrimination, thus the discrimination could potentially be improved via such an optimization process. Note, however, that the prior precision cannot be set arbitrarily high, as the modeling error of the vocal tract model increases as the smoothness assumption becomes more dominant [13].

The results showed generally comparable, but somewhat lower performance than a system based on MFCC features. However, fusion of both systems provided a substantial increase in performance compared to either of the individual systems, indicating that they offer complementary information.

Future work will investigate the performance on more realistic conditions faced in speaker verification applications. Results from previous work on pole-zero representations of nasals suggest a loss in performance when a mobile-telephone transmission channel is involved [12]. The Adaptive Multi-Rate (AMR) codec used in GSM and UMTS mobile telephone networks uses order 10 linear prediction to encode the spectral envelope, which effectively removes zeros from the spectrum. The robustness of the vocal tract estimation scheme under such conditions has not yet been investigated. Also, extensions to the vocal tract model such as paranasal cavities will be the subject of further investigations as these models provide a more realistic representation of the nasal cavity acoustics and may thus be better able to capture speaker-specific properties.

References

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

[2] T. Pruthi and C. Y. Espy-Wilson, “An MRI based Study of the Acoustic Effects of Sinus Cavities and its Application to Speaker Recognition,” Proc. Interspeech, pp. 2110–2113, 2006.

[3] J. Dang, K. Honda, and H. Suzuki, “Morphological and acoustical analysis of the nasal and paranasal cavities,” J. Acoust. Soc. Am., vol. 96(4), pp. 2088–2100, 1994.

[4] O. Fujimura, “Analysis of nasal consonants,” J. Acoust. Soc. Am., vol. 34, pp. 1865–1875, 1962.

[5] J. Glenn and N. Kleiner, “Speaker identification based on nasal phonation,” J. Acoust. Soc. Am., vol. 43, pp. 368–372, 1967.

[6] L.-S. Su, K.-P. Li, and K. S. Fu, “Identification of speakers by use of nasal coarticulation,” J. Acoust. Soc. Am., vol. 56, pp. 1876–1882, 1974.

[7] M. Sambur, “Selection of acoustic features for speaker identification,” IEEE Trans. Acoust., Speech and Sig. Proc., vol. 23, pp. 176–182, 1975.

[8] J. P. Eatock and J. S. D. Mason, “A quantitative assessment of the relative speaker discriminating properties of phonemes,” Proc. ICASSP, pp. 133–136, 1994.

[9] R. Auckenthaler, E. S. Parris, and M. J. Carey, “Improving a GMM Speaker Verification System by Phonetic Weighting,” Proc. ICASSP, pp. 313–316, 1999.

[10] B.-J. Lee, J.-Y. Choi, and H.-G. Kang, “Phonetically optimized speaker modeling for robust speaker recognition,” J. Acoust. Soc. Am., vol. 126, pp. EL100–EL106, 2009.

[11] E. Enzinger, P. Balazs, D. Marelli, and T. Becker, “A Logarithmic Based Pole-Zero Vocal Tract Model Estimation for Speaker Verification,” Proc. ICASSP, Prague, Czech Republic, pp. 4820–4823, 2011.

[12] E. Enzinger and P. Balazs, “Speaker Verification using Pole/Zero Estimates of Nasals,” Analele Universitatii “Eftimie Murgu” XVIII, 33-44, 2011.

[13] C. H. Kasess, W. Kreuzer, E. Enzinger, and N. Kerschhofer-Puhalo, “Estimation of the vocal tract shape of nasals using a Bayesian scheme,” Proc. Interspeech, 9–13 September, Portland, Oregon, U.S.A., 2012.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM,” NIST speech disc 1-1.1, NASA STI/Recon Technical Report N, 27403, 1993.

[15] K. J. Kohler (Ed.), “Phonetisch-Akustische Datenbasis des Hochdeutschen,” Kieler Arbeiten zu den PHONDAT-Projekten 1989–1992 (=Arbeitsberichte des Instituts fuer Phonetik und digitale Sprachverarbeitung der Universitaet Kiel (AIPUK) 26), 1992.

[16] I.-T. Lim and B. G. Lee, “Lossy Pole-Zero Modeling for Speech Signals,” IEEE Trans. Speech Audio Proc., vol. 4, pp. 81-88, 1996.

[17] K. Schnell, “Rohrmodelle des Sprechtraktes. Analyse, Parameterschätzung und Syntheseexperimente,” Ph.D. dissertation, Universität Frankfurt, 2000.

[18] K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny, “Variational free energy and the Laplace approximation,” Neuroimage, vol. 34, pp. 220–234, 2006.

[19] 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, 2010.

[20] K. Kalgaonkar and M. Clements, “Vocal tract and area function estimation with both lip and glottal losses,” Proc. Interspeech, pp. 550–553, 2007.

[21] 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.

[22] N. Brümmer and J. du Preez, “Application-independent evaluation of speaker detection,” Computer Speech and Language, vol. 20, pp. 230–275, 2006.

[23] D. A. van Leeuwen and N. Br¨ummer, “An introduction to application independent evaluation of speaker recognition systems,” in Speaker Classification I. Fundamentals, Features, and Methods, C. Müller, Ed., Springer, 2007, pp. 330–353.

[24] G. S. Morrison, “Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio,” Australian Journal of Forensic Sciences, vol. 45, pp. 173–197, 2013.

[25] N. Brümmer. (2005) Tools for fusion and calibration of automatic speaker detection systems. [Online]. Available: http://niko.brummer.googlepages.com/focal

[26] G. S. Morrison. (2009) Robust version of train llr fusion.m from Niko Brümmer’s FoCaL Toolbox (release 2009-07-02). [Online]. Available: http://geoff-morrison.net/#TrainFus

[27] S. Pigeon, P. Druyts, and P. Verlinde, “Applying Logistic Regression to the Fusion of the NIST’99 1-Speaker Submissions,” Digital Signal Process., vol. 10, pp. 237–248, 2000.

[28] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET curve in assessment of detection task performance,” Proc. Eurospeech, pp. 1895–1898, 1997.

[29] N. Brümmer, “Tools for ROCCH DET Curves,” https://sites.google.com/site/focaltoolkit/rocch (retrieved 14/06/2013).