BAYESIAN VOCAL TRACT MODEL ESTIMATES OF NASAL STOPS FOR SPEAKER VERIFICATION

Ewald Enzinger[1,2], *Christian H. Kasess*[1]

[1]Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria
[2]School of Elec. Eng. & Telecom., University of New South Wales, Sydney, Australia

ABSTRACT

In this paper we report on speaker verification experiments using branched vocal tract model estimates of alveolar nasal (/n/) stops. While the discriminatory potential of nasal acoustics has long been established, their acoustic properties have so far mostly been characterized using spectral features. Here, we used a Bayesian estimation technique to obtain reflection coefficients of a branched-tube model of the combined nasal and oral tract. Parameters were then modeled using probabilistic linear discriminant analysis to calculate likelihood ratios for speaker verification trials. Performance was assessed on normal and high vocal effort speech using high-quality and mobile-telephone-transmitted recordings taken from the German-language Pool2010 corpus. Results are compared with those of systems based on mel-frequency cepstral coefficients (MFCC). Vocal tract parameter based systems outperform MFCC based systems in matched conditions, but lack robustness under mismatch, while being readily interpretable with respect to a physical speech production model.

Index Terms— Nasals, vocal tract modeling, Bayesian estimation, likelihood ratio, speaker verification

1. INTRODUCTION

The acoustic properties of nasals have long been considered as an important source of speaker-discriminating information. The complicated structure of the nasal cavity 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, cause substantial acoustic variation between different speakers [1, 2]. In the production of nasal stops a closure is formed by the lips (/m/), the tongue at the alveolar ridge (/n/), or the tongue dorsum at the lowered velum (/ŋ/), while the velum is lowered, coupling the nasal cavity to the vocal tract. The relatively fixed structure of the vocal and nasal cavity provides the basis for the a-priori assumption of low within-speaker variability. Based on these theoretical aspects nasal stops have also been considered as potentially useful for performing forensic voice comparison [3, p. 133]. Early studies in speaker identification [4, 5] as well as work on the relative contribution of different sound classes and representations in automatic speaker recognition [6, 7, 8, 9, 10, 11] provided empirical evidence in support of these arguments. Recently, attempts were made to model nasal spectra using pole-zero model estimates [12, 13]. However, these studies did not explore explicit modeling of nasal acoustics. 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 more complex relation between the speech signal and the parameters 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. The present paper uses a variational Bayesian scheme to estimate the tube areas of a combined nasal and oral tract model from the log-spectrum of the speech signal of nasal stops [14]. Here, probabilistic priors are used to enforce smoothness of the tube model. Vocal tract parameters are obtained from the tube model estimates and are used as features.

Probabilistic linear discriminant analysis (PLDA) [15] is used to model these features and calculate likelihood ratios in speaker verification experiments. Evaluations are based on data from the German-language Pool2010 corpus [16]. The effect of using different prior variances in the estimation is evaluated on the basis of a development set. Performance is compared with a baseline system using mel-frequency cepstral coefficients (MFCCs) extracted from the same nasal stop segments. The effect of differences in vocal effort and mobile-telephone transmission channel are investigated.

2. METHODOLOGY

2.1. Data base

The data were extracted from recordings of 103 male adult German speakers in the Pool2010 corpus [16]. Each speaker was recorded reading a text (*The North Wind and the Sun*) using normal and high vocal effort. The latter was induced by playing 80 $\mathrm{dB}_{\mathrm{SPL}}$ white noise over headphones. In addition, a channel-degraded version of the high-quality recordings was created by transmitting them over mobile telephone. An automatic phone-level alignment [17] was performed on the recordings, followed by auditory validation of /n/ labels. The /n/ tokens in each recording were then split in equal-sized training (enrollment) and test portions. Each portion contained between 23 and 34 (median 31.5) tokens. Data of 20 speakers were used as background population for training the PLDA model parameters (see Section 2.6). Data of another 20 speakers were used as development set and data of the remaining 63 speakers as evaluation set, resulting in 63 target and 3906 non-target trials.

2.2. Vocal tract model

The details of the model have been described previously [18]. In short, the model consists of three segmented tubes connected at the velum: a pharyngeal tube (L segments), an open nasal tube (M segments), and the oral tube (N segments) which is closed at the lips. The model is parameterized in terms of a set of vocal tract reflection coefficients $\boldsymbol{\mu}$ and the nasal-oral coupling parameter σ. From this a rational transfer function $H(\boldsymbol{\mu}, \sigma, \omega) = B(\boldsymbol{\mu}, \sigma, \omega)A(\boldsymbol{\mu}, \sigma, \omega)^{-1}$

can be derived [18]. Unlike in the case of the single tube model [19], going from the $M + L + 2N$ polynomial coefficients to the $M + L + N + 1$ vocal tract parameters is in general not possible. Hence, estimation of the area function of such a model is not straight forward.

The method used to derive the model parameters is based on the estimation scheme introduced in [14]. Contrary to previous methods [18, 20] the model estimation does not rely on a separate pole-zero estimation but estimates the model from the (pre-emphasized) log-envelope $\mathbf{y}$ directly:

$$\mathbf{y}_j = \log \bar{H}\left(\boldsymbol{\theta}, \boldsymbol{\omega}_j\right) + \boldsymbol{\epsilon}_j. \tag{1}$$

For the j-th frequency $\boldsymbol{\omega}_j$ the function $\bar{H}$ evaluates the non-linear transformation from a set of vocal tract parameters $\boldsymbol{\theta}$ to the transfer function. As the reflection coefficients $\boldsymbol{\mu}$ are restricted to the open interval $(-1, 1)$ and σ is restricted to $(0, 1)$, a sigmoidal mapping from the i-th parameter $\boldsymbol{\theta}_i$ to the i-th restricted parameter is also included in $\bar{H}$ to allow for unconstrained optimization. These unrestricted parameters $\boldsymbol{\theta}$ form the VT features (VT-$\boldsymbol{\theta}$) used in the experiments. Further, a scaling factor for the transfer function is estimated which, however, is not included in the verification task. Therefore, the dimension of the parameter vector VT-$\boldsymbol{\theta}$ is $M + N + L + 1$.

2.3. Estimation scheme

The Bayesian model for the estimation scheme is given as

$$p\left(\boldsymbol{\theta}, \tau, \boldsymbol{\Pi}|\mathbf{y}\right) \propto p\left(\mathbf{y}|\boldsymbol{\theta}, \tau\right) p\left(\tau\right) p\left(\boldsymbol{\theta}|\boldsymbol{\Pi}\right) p\left(\boldsymbol{\Pi}\right), \tag{2}$$

where $\boldsymbol{\theta}$ is the vector of VT-model parameters, τ is the precision (i.e., the inverse variance) of the normally distributed likelihood $p\left(\mathbf{y}|\boldsymbol{\theta}, \tau\right) = \mathcal{N}\left(\mathbf{y}; \log \bar{H}\left(\boldsymbol{\theta}, \boldsymbol{\omega}\right), \tau \mathbf{I}\right)$, and $\boldsymbol{\Pi}$ is the precision matrix of the smoothness prior for the vocal tract parameters ($p\left(\boldsymbol{\theta}|\boldsymbol{\Pi}\right) = \mathcal{N}\left(\boldsymbol{\theta}; \mathbf{0}, \boldsymbol{\Pi}\right)$). Compared to the original scheme [14] there have been two alterations. Instead of the prior of the logarithm of τ assumed to be normal the prior of τ is now the gamma distribution ($p(\tau) = \mathrm{Gam}(\tau; a_\tau, b_\tau)$) which is the conjugate prior for the precision under the assumption of an independently and identically distributed Gaussian error. Second, previous results have shown that the choice of prior has a profound effect on the resulting parameters (e.g., [14, 21]). Thus, to avoid having to choose a fixed a-priori variance for $\boldsymbol{\theta}$, a hyperprior $p(\boldsymbol{\Pi})$ is introduced:

$$p(\boldsymbol{\Pi}) = \prod_i \mathrm{Gam}(\boldsymbol{\Pi}_i; \mathbf{a}_i, \mathbf{b}_i). \tag{3}$$

Here, the product of gamma distributions results in a diagonal prior precision matrix $\boldsymbol{\Pi}$. For a correlated prior a Wishart distribution could be used, however, here only the uncorrelated variant is used for simplicity and $\mathbf{a}_i$ and $\mathbf{b}_i$ are the same for all parameters $\boldsymbol{\theta}_i$ except the scaling which has a low precision prior. The variational estimation scheme will be described briefly. Two assumptions about the posterior density $q(\boldsymbol{\theta}, \tau, \boldsymbol{\Pi})$ are necessary. First, $q(\boldsymbol{\theta}, \tau, \boldsymbol{\Pi})$ factors as $q(\boldsymbol{\theta})q(\tau)q(\boldsymbol{\Pi})$. Second, as in the original scheme, $q(\boldsymbol{\theta})$ is assumed to be normal. Then, the posterior distributions for one set of parameters is calculated as the expected value of the log of the joint distribution $p(\boldsymbol{\theta}, \tau, \boldsymbol{\Pi})$ under the remaining two posterior distributions. The three resulting integrals are repeatedly calculated until the estimation converges. The integrals evaluating the function $\bar{H}$ are approximated using the unscented transform [22]. A Gauss-Newton scheme is used to find the mode of $q(\boldsymbol{\theta})$.

For the estimation, 30 ms long central portions of the tokens (f_s=8000 Hz) were used with a pre-emphasis of 0.9 (see also [21]). L, M, and N were chosen as 4, 6, and 4, respectively. Fundamental frequencies for the envelope extraction were estimated using [23].

2.4. Vocal tract priors

Six different settings for the $\mathbf{a}_i$ (10, 20, 50, 100, 100, 200) and $\mathbf{b}_i$ (1, 1, 1, 2, 1, 2) were evaluated. The expected value for precision is given as a/b and thus the values are 10, 20, 50, 50, 100, 100. Settings 3 and 4 as well as 5 and 6 have the same expectation respectively, however, the distribution is narrower for case 4 and 6. This range of values was chosen, as in [14] a value of 10 for the precision was found a good value for low within-subject variance and reasonable estimation error. As shown, higher values for the precision lead to less intra-subject variance which may be desirable for speaker verification. For each condition the optimal prior settings were empirically determined via tests using the development set.

2.5. Baseline features

Mel-frequency cepstral coefficients were extracted from the same 30 ms long portion of the tokens used for VT estimation. A Hanning window was applied to the non-preemphasized samples. The power spectrum was then multiplied by a filter bank consisting of 26 triangular-shaped filters with a 50% overlap. In the mel-frequency scale all filters had the same width and overlap. A discrete cosine transform (DCT) was fitted to the logarithm of the 26 filter outputs, and first 13 DCT coefficient values (MFCC values) were used as baseline features.

2.6. PLDA modeling

Vocal tract parameters (VT-$\boldsymbol{\theta}$) as well as MFCCs were directly modeled using probabilistic linear discriminant analysis (PLDA) [15]. In this approach, which is commonly used for modeling i-vector representations of recordings in automatic speaker recognition systems, the feature vectors are assumed to be generated by the generative model [15, 24, 25] (notation follows [15]):

$$x_{ij} = \mu + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij} + \epsilon_{ij}. \tag{4}$$

x_{ij} denotes the jth observation (VT-$\boldsymbol{\theta}$s or MFCCs) of speaker i, $\mu + \mathbf{F}\mathbf{h}_i$ describes the between-speaker variability, and $\mathbf{G}\mathbf{w}_{ij} + \epsilon_{ij}$ the within-speaker variability. As in [15] we use a Gaussian residual term ϵ_{ij} with diagonal covariance $\boldsymbol{\Sigma}$. The priors of the latent variables $\mathbf{h}_i$ and $\mathbf{w}_{ij}$ are assumed to be Gaussian. The model parameters μ, $\mathbf{F}$, $\mathbf{G}$, and $\boldsymbol{\Sigma}$ are trained using an Expectation-Maximization (EM) algorithm [15]. The optimal subspace dimensions N_F and N_G were empirically determined via tests using the development set[1].

2.7. Likelihood ratio calculation

Given mean vectors $\bar{x}_1$ and $\bar{x}_2$ obtained from observations of /n/ tokens in the training (enrollment) and test portions of a verification trial, a score s is calculated as a likelihood ratio with respect to two hypotheses, that both vectors share the same latent identity variable (H_1), or that they were generated from different latent identity variables (H_2):

$$s = \frac{p(\bar{x}_1, \bar{x}_2|H_1)}{p(\bar{x}_1|H_2)p(\bar{x}_2|H_2)}. \tag{5}$$

[1] An initial approach using GMM-UBM [26] was discarded based on inferior performance on tests using the development set.

Scores obtained from tests on the development set were used to calculate weights for logistic-regression calibration [27, 28] which was applied to calibrate the scores from the evaluation set. Logistic regression was also used to fuse the scores from VT-$\boldsymbol{\theta}$ and MFCC based systems [29].

3. RESULTS

We assessed system performance using Equal Error Rates (EER) and the log likelihood ratio cost (C_{llr}) metric, as well as Detection error trade-off (DET) plots as graphical representations. EER and DET plot statistics were obtained using the Receiver Operator Characteristic Convex Hull method[2].

3.1. Vocal tract prior settings

We first investigated the performance of the six different prior settings via tests using normal vocal effort speech and high-quality recordings of the development set. The results in Table 1 suggest that higher values for the precision lead to better speaker verification performance on the development set.

	1	2	3	4	5	6
EER	2.90	3.26	3.62	4.00	2.78	1.52
C_{llr}	0.180	0.185	0.179	0.181	0.108	0.082

Table 1. EER and C_{llr} of tests of VT-$\boldsymbol{\theta}$ based systems using different VT prior settings (see Section 2.4).

3.2. Normal vocal effort, high-quality recordings

Figure 1 shows the results of tests on normal vocal effort speech from high-quality recordings. The VT-$\boldsymbol{\theta}$ based system shows better performance than the MFCC based system, except for operating points in the low false alarm probability region (due to skewed, non-Gaussian score distributions). Fusion of both systems increases performance over both individual systems.

3.3. High vocal effort, high-quality recordings

Figure 2 and Table 2 show the results for tests on high vocal effort speech using high-quality recordings. Solid lines indicate performance on tests with matched conditions, i.e., both comparison samples have high vocal effort. Dashed lines show performance on tests with mismatched conditions, i.e., one sample has high vocal effort, the other normal vocal effort. In tests on matched conditions the VT-$\boldsymbol{\theta}$ based system shows better performance than the MFCC based system. Under mismatch the general system performance decreases substantially, in particular that of the VT-$\boldsymbol{\theta}$ based system.

high vocal effort	Matched		Mismatched	
	EER	$\mathbf{C_{llr}}$	**EER**	$\mathbf{C_{llr}}$
VT-$\boldsymbol{\theta}$ ($N_F = N_G = 10$)	3.30	0.317	15.00	0.574
MFCC ($N_F = N_G = 10$)	4.20	0.201	11.30	0.405
Fusion	1.50	0.133	9.80	0.333

Table 2. EER and C_{llr} of VT-$\boldsymbol{\theta}$ and MFCC based systems (high v normal vocal effort, high-quality recordings)

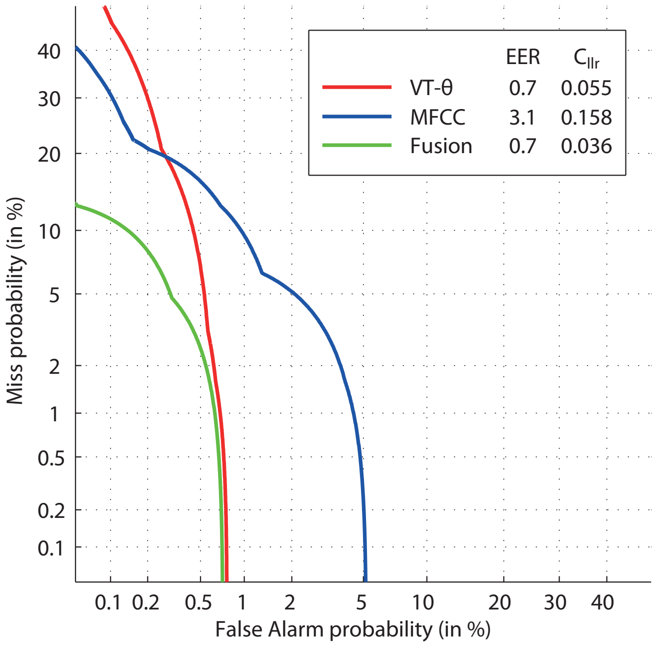

Fig. 1. DET plot comparing the performance of alveolar nasal stop (/n/) VT-$\boldsymbol{\theta}$ and MFCC based systems (normal vocal effort, high-quality recordings).

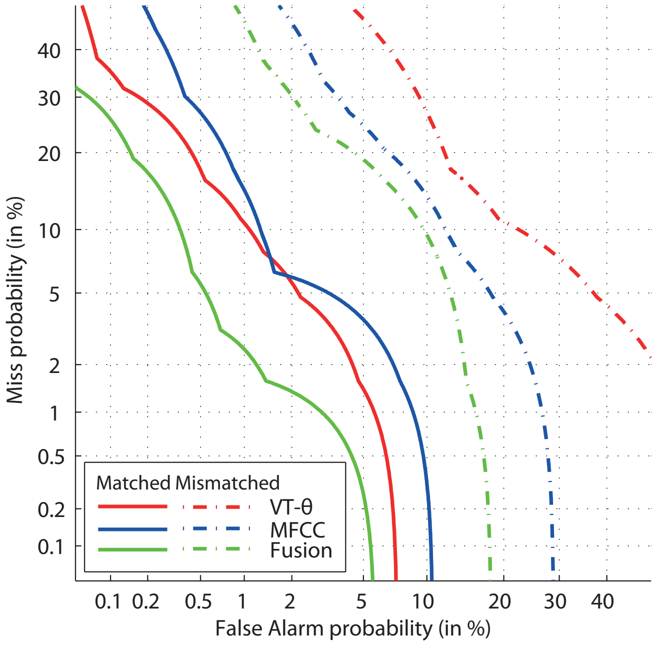

Fig. 2. DET plot comparing the performance of alveolar nasal stop (/n/) VT-$\boldsymbol{\theta}$ and MFCC based systems (high v normal vocal effort, high-quality recordings).

3.4. Normal vocal effort, mobile-telephone channel

Figure 3 and Table 3 show the results for tests on mobile-telephone channel using normal vocal effort speech. Solid lines indicate per-

[2] `http://focaltoolkit.googlepages.com/rocch`

formance on tests with matched conditions, i.e., both comparison samples are from mobile-telephone recordings. Dashed lines show performance on tests with mismatched conditions, i.e., one sample is from a mobile-telephone recordings, the other from a high-quality recording. As with vocal effort, the VT-$\boldsymbol{\theta}$ based system shows better performance than the MFCC based system in matched conditions, but worse performance in mismatched conditions.

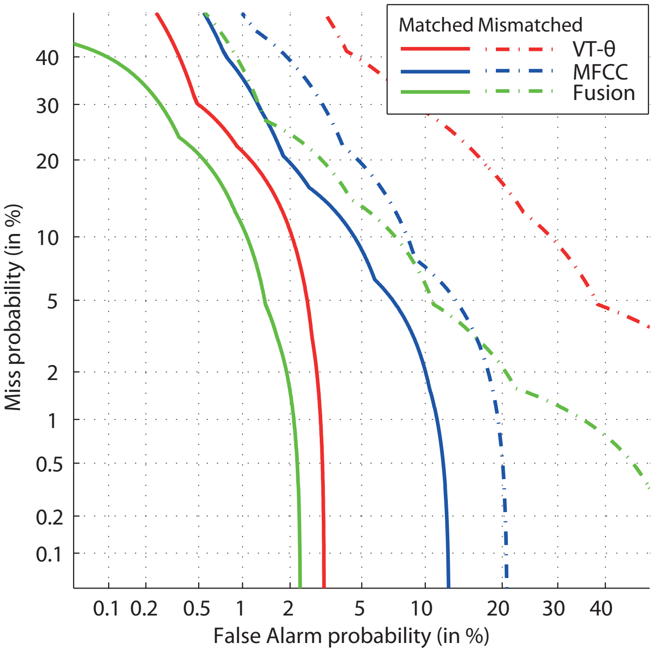

Fig. 3. DET plot comparing the performance of alveolar nasal stop (/n/) VT-$\boldsymbol{\theta}$ and MFCC based systems (normal vocal effort, mobile-telephone v high-quality).

mobile-telephone	Matched		Mismatched	
	EER	$\mathbf{C_{llr}}$	**EER**	$\mathbf{C_{llr}}$
VT-$\boldsymbol{\theta}$ ($N_F = N_G = 11$)	2.70	0.155	18.30	1.265
MFCC ($N_F = N_G = 8$)	6.10	0.229	8.80	0.401
Fusion	1.90	0.102	8.40	0.686

Table 3. EER and C_{llr} of VT-$\boldsymbol{\theta}$ and MFCC based systems (normal vocal effort, mobile-telephone v high-quality)

4. DISCUSSION AND CONCLUSION

The present paper assesses the performance of physiologically motivated vocal tract model estimates of alveolar nasal stop (/n/) tokens in speaker verification experiments. Parameters of a branched-tube model of the combined nasal and oral tract are obtained using a Bayesian estimation technique. These are then modeled using probabilistic linear discriminant analysis.

As noted in Section 2.4, higher values for the precision in the Bayesian vocal tract estimation generally lead to less intra-subject variance. Correspondingly, we observed that performance on the development set increased with higher precision values. In tests where the levels of vocal effort or the transmission channel conditions were matched between the comparison samples, performance of VT-$\boldsymbol{\theta}$ based systems compared favorably to that of MFCC based systems. Fusion of both systems generally improved upon both individual systems, indicating that they offer complementary information. However, results under mismatched conditions indicate that the VT-$\boldsymbol{\theta}$ based systems are less robust.

The VT model estimation is based on an estimate of the log-spectral envelope. Differences in fundamental frequency (f0) may have a profound effect on this estimate. Studies on the effect of high vocal effort on f0 found an increase in average f0 with higher vocal effort [16]. With respect to mobile-telephone channel, 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 may affect the vocal tract estimation. Future work will focus on mitigating those effects. Also, extensions to the vocal tract model such as paranasal cavities [21] 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.

5. REFERENCES

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

[2] Gunnar Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.

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

[4] J.W. Glenn and N. Kleiner, “Speaker Identification Based on Nasal Phonation,” *J. Acoust. Soc. Am.*, vol. 43, no. 2, pp. 368–372, 1967.

[5] L.-S. Su, K.-P. Li, and K. S. Fu, “Identification of speakers by use of nasal coarticulation,” *J. Acoust. Soc. Am.*, vol. 56, no. 6, pp. 1876–1882, 1974.

[6] M. Sambur, “Selection of acoustic features for speaker identification,” *IEEE Trans. Acoust., Speech and Sig. Proc.*, vol. 23, no. 2, pp. 176–182, 1975.

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

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

[9] B.-J. Lee, J.-Y. Choi, and H.-G. Kang, “Phonetically optimized speaker modeling for robust speaker recognition,” *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. EL100–EL106, 2009.

[10] N. Scheffer, L. Ferrer, M. Graciarena, S. Kajarekar, E. Shriberg, and A. Stolcke, “The SRI NIST 2010 speaker recognition evaluation system,” in *Proc. ICASSP*, 2011, pp. 5292–5295.

[11] H. Lei and E. López-Gonzalo, “Importance of nasality measures for speaker recognition data selection and performance prediction,” in *Proc. Interspeech*, 2009, pp. 888–891.

[12] E. Enzinger, P. Balazs, D. Marelli, and T. Becker, “A logarithmic based pole-zero vocal tract model estimation for speaker verification,” in *Proc. ICASSP*, 2011, pp. 4820–4823.

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

[14] C. H. Kasess, W. Kreuzer, E. Enzinger, and N. Kerschhofer-Puhalo, "Estimation of the vocal tract shape of nasals using a Bayesian scheme," in *Proc. Interspeech*, 2012.

[15] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.

[16] 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.

[17] 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.

[18] I.-T. Lim and B.G. Lee, "Lossy pole-zero modeling for speech signals," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 2, 1996.

[19] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-21, no. 5, pp. 417–427, 1972.

[20] K Schnell, *Rohrmodelle des Sprechtraktes. Analyse, Parameterschätzung und Syntheseexperimente*, Ph.D. thesis, Goethe University Frankfurt am Main, 2003.

[21] C.H. Kasess and W. Kreuzer, "Estimation of multiple-branch vocal tract models: the influence of prior assumptions," in *Proc. Interspeech*, 2013, pp. 1663–1667.

[22] S. Julier and J. Uhlmann, "New extension of the Kalman filter to nonlinear systems," *Proceedings of the 1997 SPIE Conference on Signal Processing, Sensor Fusion, and Target Recognition*, vol. 3068, pp. 182–193, 1997.

[23] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech*, 2011, pp. 1973–1976.

[24] Patrick Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2010.

[25] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.

[26] 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.

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

[28] D. A. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification I. Fundamentals, Features, and Methods*, C. Müller, Ed., pp. 330–353. Springer, 2007.

[29] Stéphane Pigeon, Pascal Druyts, and Patrick Verlinde, "Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions," *Digital Signal Process.*, vol. 10, pp. 237–248, 2000.