ANALELE UNIVERSITĂŢII
"EFTIMIE MURGU" REŞIŢA
ANUL XVIII, NR. 1, 2011, ISSN 1453 - 7397

Ewald Enzinger, Peter Balazs

Speaker Verification using Pole/Zero Estimates of Nasals

The acoustics of nasals are an important source of speaker-discriminating features. Nasal spectra contain poles and zeros dependent upon nasal cavities which are complex static structures which vary from person to person. Nasal spectra may therefore have low within-speaker and high between-speaker variability. This study applies a recent pole-zero model estimation technique based on a logarithmic criterion on nasal spectra to obtain pole/zero features for speaker verification. The robustness against two mismatch conditions, Lombard speech and studio versus GSM transmission channel, is evaluated and compared with an approach based on MFCC features. Furthermore, results of fusion of the nasal systems with a generic MFCC-based GMM-UBM speaker verification system are presented.

Keywords*: Pole/Zero model, nasals, speaker verification*

1. Introduction

For an acoustic property to be useful for discriminating speakers, its between-speaker variability has to be greater than its within-speaker variability. During the production of nasal stops, the lowered velum couples the nasal cavities with the vocal tract 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/). During the oral closure, the articulators essentially don't move, which is reflected in the relatively stationary spectrogram of the output sound pressure wave radiated from the nostrils. The relatively fixed posture of the vocal and nasal cavities provides the basis for the a-priori assumption of low within-speaker variability.

Following the source-filter theory [1], the glottal pulse stream provides the excitation for the vocal tract which shapes the spectrum of the sound pressure wave due to its resonances. In nasals, the pathway created by the pharynx and the nasal cavity 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 with 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 [4,1]. Combined with the property of low within-speaker variability, spectra of nasal stops are theoretically valuable sources of speaker-discriminating features.

The significance of explicit use of nasal segments was demonstrated in early studies on speaker identification [5,6]. In the domain of 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].

In this paper we evaluate representations based on pole/zero model estimates based on a logarithmic criterion [11] on different mismatch conditions. This work represents an extension of our previous study [12]. In Section 2 we give a description of the pole/zero model estimation method. The speaker verification approach and the evaluation design chosen for this work are presented in Sections 3 and 4. We present the results including the effects of different mismatch conditions, compared with MFCC features as baseline, in Section 5, and conclude with a discussion of these results in Section 6.

The task of estimating the poles and zeroes of the vocal / nasal tract can be seen as parallel to a modal analysis of the physical system of the vocal tract. In this paper we use the method in [11] which is based on a purely signal processing approach. This system identification approach is one step closer to the physical model compared to MFCC features, where the link to the physiological properties is unclear. In [23] it is shown that this pole/zero model can be linked to a two tube model of the vocal / nasal tract.

In addition to having the conceptual advantages of being closer realted to the physical properties, we show in this paper that the proposed features outperforms or matches the performance of the standard MFCC features on nasal stop segments.

2. Pole/Zero model estimation

From a signal processing perspective, speech production can be modeled by a linear, slowly time-varying filter which incorporates the combined effect of the vocal and/or nasal tract and the radiation at the lips or nostrils as well as the glottal pulse shape. The sampled speech signal is generated by an excitation signal, assumed to be a train of impulses for voiced sounds, which is filtered by the speech production filter which is assumed to be LTI. The frequency response of this filter

is given in equation (1), which represents the pole-zero model, where *n* and *m* denote the orders of the numerator and denominator, respectively.

$$G(z) = \frac{B(z)}{A(z)} = \frac{\sum_{l=0}^{n} b^l z^{-l}}{\sum_{l=0}^{m} a^l z^{-l}} \quad (1)$$

The estimation of the parameters of pole-zero models from speech has been thoroughly studied in the field of digital signal processing. Different approaches have been proposed which differ with respect to their robustness towards noise and other detrimental conditions present in the speech data. We use a recently published method [11] which we previously employed in obtaining positions of poles and zeros as features for speaker verification [12].

Motivated by the perception of amplitude by the human auditory system, the set of parameters θ consisting of the coefficients of the numerator and denominator polynomials of the pole/zero model is optimized by minimizing a logarithmic criterion (equation 2).

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

$\hat{G}(\omega_k)$ is the frequency response estimate at discrete frequencies $\omega_{k=1,\ldots,K}$ obtained by interpolating spectral peaks found within neighborhoods of multiples of f_0. An initial estimate of the coefficients is obtained from a minimum-phase estimate of the frequency response by a weighted linear least-squares (WLLS) algorithm [13]. The method optimizes the numerator and denominator coefficients directly in an iterative procedure. The details of the algorithm are described in [11,12]. Figure 1 shows an example of pole/zero model estimation obtained from an alveolar nasal stop (/n/).

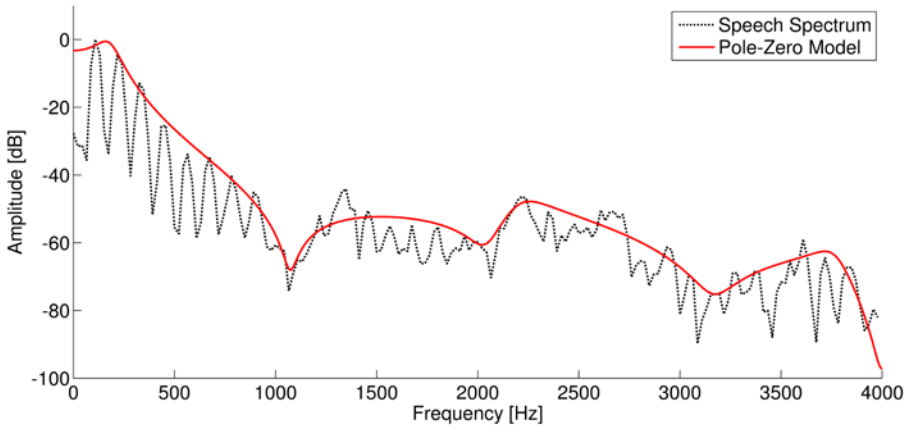

Figure 1. Pole/zero model estimate of an alveolar nasal stop (/n/).

3. Speaker verification system

The Gaussian mixture model – universal background model (GMM-UBM) approach [14] is adopted in this work. We favored this technique over current state-of-the-art speaker verification approaches such as GMM-SVM, JFA and i-vector based systems, as the benefits of these systems depend to a large extend on additional training data, which is often not easily available, e.g. in forensic applications. This work aims at evaluating the robustness of pole/zero estimates for speaker verification in the feature domain and does not employ channel or session compensation methods in the score domain.

The pole-zero model based features extracted from the nasal segments are given by the pole and zero frequencies, i.e. the angular positions of the locations of roots of the numerator and denominator polynomials evaluated in the frequency domain. The features are extracted using a shifted 30 ms Hamming window with 90% overlap. No further pre-processing such as pre-emphasis is applied. An order of 11 was used for both the numerator and denominator polynomials in the model estimation. The locations of the roots of the polynomials were sorted in ascending order and the first three poles and zeros were selected as features, resulting in 6 features per frame. No additional tracking procedures such as described in [15], which are commonly applied in formant tracking, were employed.

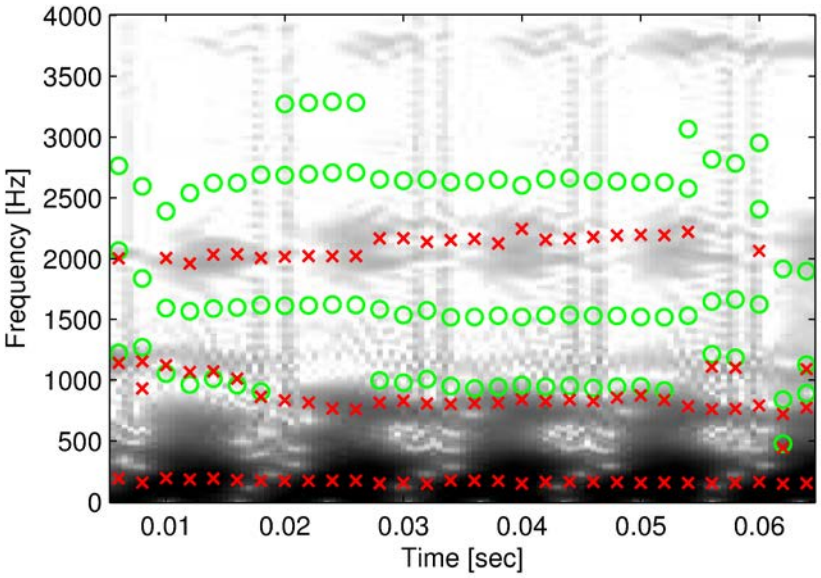

Figure 2. Pole/zero frequencies estimated in an alveolar nasal stop (/n/).

Feature vectors are modeled by mixtures of Gaussian distributions (GMMs) with full covariance matrices, denoted by

$$\lambda := (p_i, \mu_i, \Sigma_i)_{i=1,\ldots,M} \tag{3}$$

where p_i, μ_i and Σ_i represent the mixture weights, means and covariance matrices. The universal background model (UBM), which models the distribution of the features in the reference population, is trained on the background data pooled across speakers. Its mixture weights, means and co-variances are estimated using the expectation-maximization (EM) algorithm. Actual speaker GMMs are generated by maximum a-posteriori (MAP) adaption of the UBM means.

4. Evaluation

The evaluation is based on data of 103 male German speakers from the Pool2010 corpus [16]. It contains recordings of read as well as spontaneous speech, both in a normal speaking style condition and a Lombard condition where the speakers were subjected to background noise played back over headphones. The Lombard effect is characterized by increased vocal effort while speaking in a noisy environment. It is highly variable between speakers, as they use different strategies for increasing intensity from normal to Lombard speech [17]. The speech data was recorded to disk as well as transmitted via a GSM channel. These recordings are used to test the robustness to transmission channel and speaking-style mismatch.

The procedure outlined in [12] was adopted to obtain segmentations of two nasal consonants /n/ and /m/ from the read speech portion of the dataset by performing automatic phone-level alignment. The resulting segments were validated by listening and selecting only correctly labeled tokens. The features from nasal spectra were evaluated by splitting the recordings and comparing pole/zero estimates from nasals in the first half of one condition against the segments in the second half of the other condition.

The nasal stops /n/ and /m/ were modeled and evaluated separately. The data was split into a background set of 20 speakers for UBM training, a development set of 20 speakers, and a test set of the remaining 63 speakers. Scores were calculated for each pair of speakers in the development set data, resulting in 20 target and 380 non-target trials. These were then used to calculate weights for logistic-regression calibration and fusion [18,19,20]. The optimal number of Gaussian mixtures and iterations for MAP adaptation was determined based on evaluation of the trials of the development set using the log-likelihood ratio cost (C_{llr}) metric [19].

Evaluations were performed on the test set using the same procedure, using the optimal number of Gaussian mixtures and iterations of MAP adaptation obtained from the development set. The resulting scores from the test set were then

calibrated to log-likelihood ratios using the weights which have been calculated using the scores form the development set.

As in the original approach [12], we use Mel frequency cepstral coefficients (MFCCs) as a baseline for comparison. 13 MFCCs are extracted using a shifted 30 ms hamming window with 90% overlap. We used a configuration similar to HTK, i.e. 20 Mel filters, a pre-emphasis factor of 0.97 and -22 as the liftering exponent.

Finally, the results from the individual nasal systems were fused with a generic MFCC-based GMM-UBM automatic speaker verification system. 13 Mel frequency cepstral coefficients (MFCCs) were extracted every 10 ms from the speech portion of each recording using a 20 ms hamming window. The feature vectors were modeled by Gaussian mixture models with diagonal covariance matrices using 1024 mixture components. MAP adaptation was used to create individual speaker models from the UBM. The system follows the same procedure as the nasal systems for calibration and fusion of the scores resulting from the test dataset, using the weights obtained from the development set to combine the generic automatic system with both individual nasal systems.

5. Results

The evaluation results are presented in terms of the log-likelihood ratio cost (C_{llr}) metric as well as equal error rates (EER) and detection error trade-off plots obtained from the ROCCH procedure [21].

First, the performance of the pole/zero features is assessed and compared with MFCC features for both nasal stops. Table 1 shows the baseline performance on same-session data without mismatch. The pole/zero features outperform the MFCC features both in terms of equal error rate and log-likelihood ratio cost, for both nasal stops individually as well as when fused together. Figure 3 shows the respective DET curves.

Table 1. Performance of pole/zero and MFCC features for nasal stops.

	Pole/zero frequencies		MFCC	
	C_{llr}	EER	C_{llr}	EER
/m/	0.424	12.6%	0.571	14.5%
/n/	0.153	3.7%	0.246	5.8%
fused	0.150	3.6%	0.250	5.9%

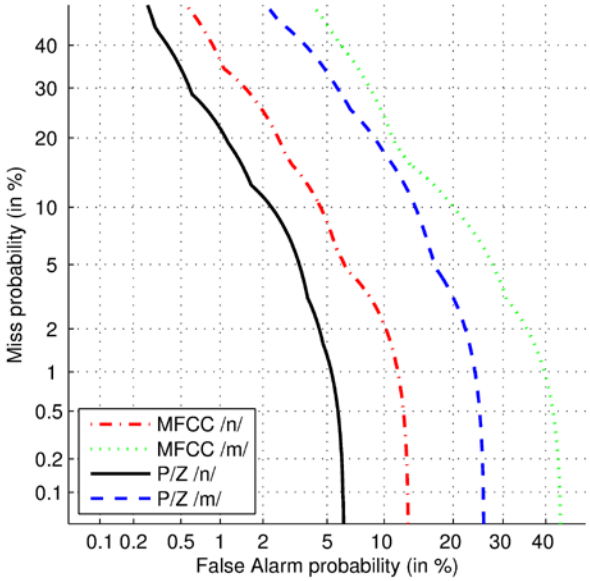

Figure 3. DET plot of pole/zero and MFCC features on both nasal segments.

In a comparison of trials with mismatch in vocal effort between free speech and Lombard condition, the performance of both features decreases, with MFCC displaying slightly lower EER and C_{llr} values (Table 2).

Table 2. Performance of pole/zero and MFCC features in free vs. Lombard condition mismatch.

	Pole/zero frequencies		MFCC	
	C_{llr}	EER	C_{llr}	EER
/m/	0.708	22.5%	0.732	21.3%
/n/	0.560	12.6%	0.528	13.1%
fused	0.547	12.9%	0.504	12.2%

The mismatch in transmission channel, especially for GSM due to its inherent all-pole modeling of speech, is expected to affect the performance of pole/zero features. The evaluation results are given in Table 3.

Table 3. Performance of pole/zero and MFCC features in studio vs. GSM transmission channel mismatch.

	Pole/zero frequencies		MFCC	
	C_{llr}	EER	C_{llr}	EER
/m/	0.949	36.4%	0.988	45.4%
/n/	1.109	40.4%	0.979	42.4%
fused	1.054	39.8%	0.975	40.4%

As can be seen, both features show very high error rates and C_{llr} values near or above unity. The EER values for the pole/zero feature based systems are lower than results from the MFCC based approach.

Evaluation results for trials featuring both mismatch conditions combined are given in Table 4. As in the previous mismatch condition, the system performance is strongly decreased for both methods, with pole/zero features again providing somewhat lower EER values than MFCC features.

Table 4. Performance of pole/zero and MFCC features in combined GSM transmission channel and Lombard condition mismatch.

	Pole/zero frequencies		MFCC	
	C_{llr}	EER	C_{llr}	EER
/m/	0.969	39.3%	1.053	49.6%
/n/	0.982	39.2%	1.181	45.0%
fused	0.966	39.1%	1.252	46.2%

Finally, the pole/zero systems for both nasal stops /m/ and /n/ are fused with a generic automatic MFCC-based GMM-UBM speaker verification system which operates on the whole speech portion of the recordings, indiscriminately including the nasal segments, using fusion weights determined from the development set. Table 5 shows the performance of the speaker verification system by itself as well as after fusion for both mismatch conditions as well as both combined.

Table 5. Performance of pole/zero features fused with generic automatic MFCC based GMM-UBM speaker verification system under mismatch.

	Generic system		Fused Pole/Zero /n/+/m/	
	C_{llr}	EER	C_{llr}	EER
GSM	0.022	0.4%	0.024	0.3%
Lombard	0.492	9.7%	0.371	8.5%
combined	0.461	11.2%	0.483	11.0%

A relatively small, but consistent increase in performance in terms of EER over the system baseline can be observed. Except for the mismatch between free speech and Lambert condition, C_{llr} values are not improved. A DET plot comparing the baseline and fused systems is given in Figure 4.

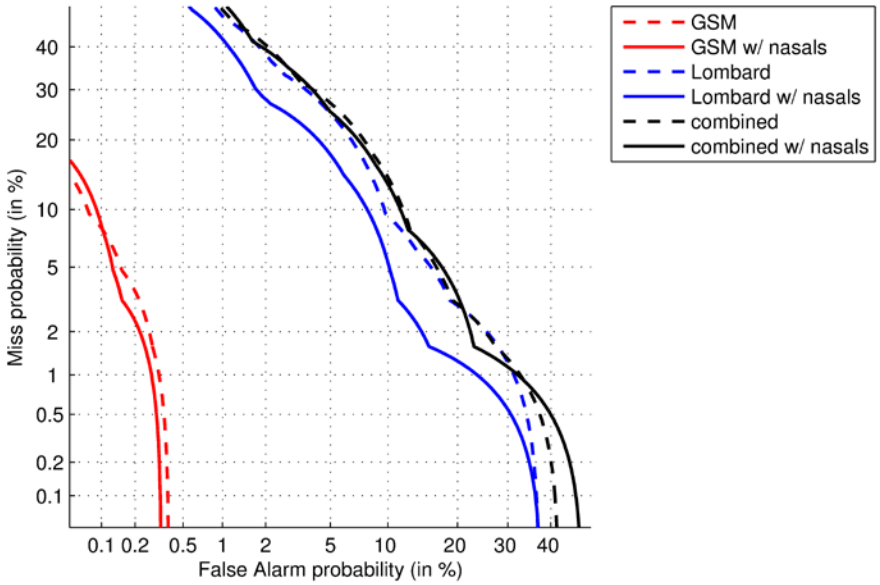

Figure 4. Generic automatic MFCC-based GMM-UBM speaker verification system fused with both Pole/Zero feature based nasal systems for mismatch in transmission channel, Lombard condition and both mismatch conditions combined.

6. Discussion and conclusion

In this paper, a new set of features consisting of pole and zero frequencies, i.e. the angular positions of the locations of roots of the numerator and denominator polynomials evaluated in the frequency domain, are obtained using a logarithmic based pole-zero model estimate of the speech production filter [11]. Theoretically, these features are advantageous for modeling the acoustics of nasal spectra and therefore provide a more straightforward interpretation with respect to models of the vocal and nasal tract [22,23].

The resulting features are evaluated in a speaker verification task in two different mismatch conditions as well as both mismatch conditions combined. The proposed pole/zero features consistently outperform or match the performance of MFCC features on nasal stop segments. Compared to MFCCs, the dimensionality of the feature vectors is greatly reduced (6 pole/zero frequencies compared to 12-16 MFCCs). This property is useful with respect to GMM training on a low number of nasal tokens.

The loss in performance for the transmission channel mismatch condition can most readily be explained by the fact that 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 spec-

trum. Due to this inherent all-pole modeling, the resulting estimates for the pole/zero frequencies are therefore expected to be greatly distorted. However, the MFCC features showed to be equally affected. Further tests using channel compensation techniques are needed in order to assess whether the robustness of the pole/zero features to transmission channel mismatch can be improved.

The fusion of pole/zero features with a generic automatic speaker verification system showed consistent, but only rather minor improvements in performance. These results should be taken as preliminary, as the automatic system used the whole speech portion of the recordings, which contain the same text read by every speaker. Further tests are required based on a more diverse and realistic database.

Future work will investigate the use of a perceptual frequency scale (e.g. Bark) for pole/zero model estimation, as it is applied in Perceptual Linear Prediction (PLP). Furthermore, procedures for tracking poles and zeros to obtain formant and antiformant tracks [15] can be used to improve the robustness of the features.

References

[1] Fant, G., *Acoustic theory of speech production*. The Hague: Mouton, 1960.

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

[3] Dang, J., Honda, K., Suzuki, H., *Morphological and acoustical analysis of the nasal and paranasal cavities*. J. Acoust. Soc. Am. 96/4, 2088-2100, 1994.

[4] Fujimura, O., *Analysis of nasal consonants*. J. Acoust. Soc. Am. 34, 1865-1875, 1962.

[5] Glenn, J., Kleiner, N., *Speaker identification based on nasal phonation*. J. Acoust. Soc. Am. 43, 368-372, 1967.

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

[7] Sambur, M., *Selection of acoustic features for speaker identification*, IEEE Trans. Acoust., Speech and Sig. Proc. 23, 176-182, 1975.

[8] Eatock, J.P., Mason, J.S.D., *A quantitative assessment of the relative speaker discriminating properties of phonemes*, Proc. ICASSP, 133-136, 1994.

[9] Auckenthaler, R., Parris, E.S., Carey, M.J., *Improving a GMM Speaker Verification System by Phonetic Weighting*, Proc. ICASSP, 313-316, 1999.

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

[11] Marelli, D., Balazs, P. *On Pole-Zero Model Estimation Methods Minimizing a Logarithmic Criterion for Speech Analysis*. IEEE Trans. Audio Speech Lang. Process. 18/2, 237-248, 2010.
[12] Enzinger, E., Balazs, P., Marelli, D., Becker, T., *A logarithmic based pole-zero vocal tract model estimation for speaker verification*. Proc. ICASSP, 4820-4823, Prague, Czech Republic, 2011.
[13] Kobayashi, T., Imai, S., *Design of IIR digital filters with arbitrary log magnitude function by WLS techniques*. IEEE Trans. Acoust., Speech, Signal Process. 38/2, 247–252, 1990.
[14] Reynolds, D.A., Quatieri, T.F., Dunn, R.B., *Speaker verification using adapted Gaussian mixture models*, Digit. Signal Process. 10, 19-41, 2000.
[15] Mehta, D.D., Rudoy, D., Wolfe, P.J., *KARMA: Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking*, arXiv:1107.0076v1.
[16] Jessen, M., Köster, O., Gfroerer, S., *Influence of vocal effort on average and variability of fundamental frequency*, Int. J. Speech, Language, and the Law 12/2, 174-213, 2005.
[17] Junqua, J.-C., *The Lombard reflex and its role on human listeners and automatic speech recognizers*. J. Acoust. Soc. Am. 93/1, 510-524, 1993.
[18] Brümmer, N., *Tools for fusion and calibration of automatic speaker detection systems*, URL http://niko.brummer.googlepages.com/focal, 2005.
[19] Brümmer, N., du Preez, J., *Application independent evaluation of speaker detection*, Comput. Speech Lang. 20, 230-275, 2006.
[20] Brümmer, N., Burget, L., Černocký, J., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D.A., Matějka, P., Schwarz, P., Strasheim, A., *Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006*, IEEE Trans. Audio Speech Lang. Proc. 15, 2072-2084, 2007.
[21] Brümmer, N., *Measuring, refining and calibrating speaker and language information extracted from speech*. Dissertation. University of Stellenbosch, 2010.
[22] Lim, I.-T., Lee, B.G., Lossy pole-zero modelling for speech signals, IEEE Trans. Speech Audio Proc. 4, 81-88, 1996.
[23] Kreuzer, W., Balazs, P., Marelli, D., *Vokaltraktmodellierung unter Verwendung eines Pol-Nullstellen Modells*, Proc. DAGA., 283-284, Düsseldorf, Germany, 2011.

Addresses:

- Ewald Enzinger, Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12-14, 1040 Wien, Austria; Forensic Voice Comparison lab, School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, Australia, ewald.enzinger@oeaw.ac.at
- Dr. Peter Balazs, Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12-14, 1040 Wien, Austria, peter.balazs@oeaw.ac.at