

Lombard Speech Recognition: A Comparative Study

H. Bořil¹, P. Fousek¹, D. Sündermann², P. Červa³, J. Žďánský³

¹*Czech Technical University in Prague, Czech Republic
{borilh, p.fousek}@gmail.com*

²*Siemens Corporate Technology, Munich, Germany
david@suendermann.com*

³*Technical University of Liberec, Czech Republic
{petr.cerva, jindrich.zdansky}@tul.cz*

Abstract: In this paper, new approaches to recognition of speech under Lombard effect are proposed. Performances of front ends with modified filter banks, multi-resolution RASTA features, voice conversion and acoustic models adaptation to talking style and speaker are evaluated and compared to the common recognition schemes. The presented experiments are carried out on Czech SPEECON and CLSD'05 databases. It is shown that under Lombard effect, all proposed approaches outperform the standard HMM system with MFCC or PLP features.

1. Introduction

One of the major tasks in recent ASR is the design of a recognition system operating reliably in real environments. Many past studies have reported a severe deterioration of recognition performance when dealing with speech uttered in adverse noisy conditions. In the last decades, major effort has been dedicated to the development of noise suppression, speech enhancement and model adaptation algorithms. However, it has been observed that even if the noise background in the speech signal is reduced, speech production variations introduced by a speaker in an effort to preserve intelligibility in adverse conditions (Lombard effect – LE) may further significantly corrupt the ASR performance [1].

In this paper, several novel approaches to speech recognition are compared to the commonly used schemes under neutral and LE conditions. At first, front ends with modified filter banks, multi-resolution RASTA features and voice conversion are described and tested in the digits recognition task. Secondly, the large vocabulary continuous speech recognition (LVCSR) system is introduced and evaluated on three types of speech: neutral, LE and LE normalized to neutral by means of voice conversion (VC). Thirdly, the performance of acoustic models adaptation to talking style and speaker is examined on the digits and LVCSR tasks.

All presented experiments are carried out on Czech SPEECON [2] and CLSD'05 (Czech Lombard Speech Database) [3] databases.

2. Front ends for robust Lombard speech recognition

The front ends presented in this section were tested in HMM-based recognition systems. The models were trained on 37 SPEECON female office sessions (37 speakers, approximately 10 hours of signal). The recordings were down-sampled from 16 kHz to 8 kHz and filtered by the telephone filter G.712. Digits from four neutral and LE CLSD'05 female sessions formed the open test set, which was used for evaluation.

2.1. Front ends with modified filter banks

Front ends with modified filter banks were designed using a data-driven approach. 8 Neutral and 8 LE CLSD'05 female sessions formed the development set (disjunctive from the open test set). The triangular filter bank (FB) used for deriving the well-known mel frequency cepstral coefficients (MFCC) [4] was replaced by various configurations of FBs with rectangular filters distributed over the linear frequency scale. The rest of MFCC algorithm was kept.

It was found that a bank of 20 rectangular filters of the constant width distributed without overlap in the interval 0-4 kHz (linear frequency cepstral coefficients – LFCC, 20 bands) provides better performance on LE speech, when compared to the former triangular mel FB, and a similar performance on neutral speech. Furthermore, omitting low frequency components by placing 19 rectangular filters from 645 Hz to 4 kHz (LFCC, 19 bands) significantly increased the accuracy on LE speech while slightly decreasing the performance on neutral speech. Finally, an algorithm iteratively repartitioning the 19-band FB was applied yielding further improvement (repartitioned frequency cepstral coefficients – RFCC) [5].

2.2. Multi-resolution RASTA features

Multi-resolution RASTA features (MR-ANN) [6] are extracted in 2 stages. Firstly, an auditory spectrum with 15 bands is calculated from the speech as in PLP analysis [14]. The time trajectory of these sub-band energies is filtered with a bank of two-dimensional filters, yielding a set of about 500 coefficients every 10 ms. In the second step, an artificial neural network (ANN) projects the coefficients to posterior probabilities of phonemes, reducing the feature size. The posteriors are then decorrelated and gaussianized using logarithm and principal component analysis in order to better fit the subsequent GMM/HMM model, see Fig. 1.

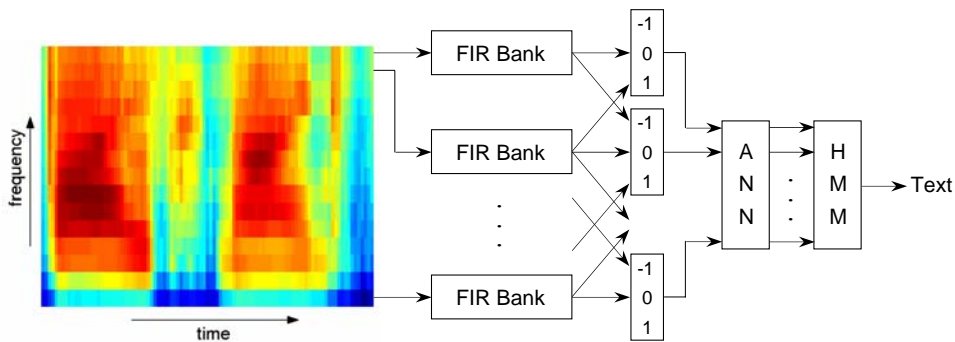


Figure 1: Speech recognition using multi-resolution RASTA features.

2.3. Voice conversion

Voice conversion (VC) is a transformation of the source voice parameters towards a target voice [7]. Different properties of the voice, namely characteristics of the vocal tract (formant positions, bandwidths), excitation and prosody (behavior of the fundamental frequency) are usually addressed separately in the VC. Since LE affects all these characteristics, VC may be used as a tool normalizing LE speech towards neutral speech. Such a converted speech may better fit models that were trained on neutral utterances.

The present work uses a VC system based on linear transformation as described in [8], see Fig. 1. Here, the source speech is Lombard-affected, and the target is its ‘neutralized’ version with respect to the fundamental frequency and formant locations. The excitation is

preserved in this case. The speech to be converted is split into pitch-synchronous¹ frames by means of the automatic pitch tracking described in [9].

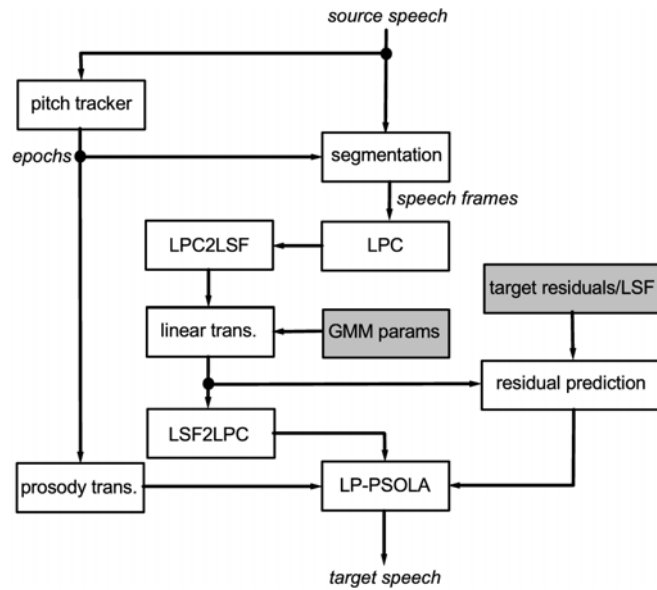


Figure 2: Components of the voice conversion system used in this study.

These frames are featurized to line spectral frequencies (LSF) showing superior manipulation properties compared with other features as LPCs or MFCCs [10]. Subsequently, a linear transformation is applied to the LSF features, intending to change the vocal tract characteristics towards the target voice. The parameters of the transformation are estimated in the training phase using the approach described in [11].

The converted features are transformed back to time domain using a procedure called residual prediction [12] and concatenated by means of pitch-synchronous overlap and add [13], where the fundamental frequency is changed towards the target voice. The resulting speech that better resembles the target voice characteristics substitutes the original speech at the input of the speech recognizer.

The presented investigation exploited the fact that parallel utterances from each speaker were available, i.e. pairs of the same utterance with and without Lombard effect that allowed for properly training the transformation parameters. Mean fundamental frequency of source and target were derived from the mean frame lengths.

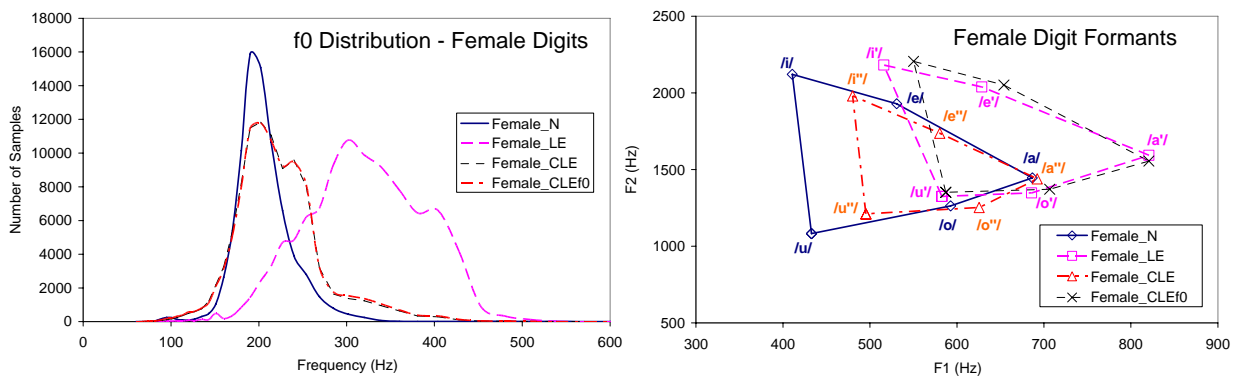


Figure 3: Distribution of F_0 in female digits and average locations of F_1 and F_2 in their vowels.

¹ In speech synthesis and related technologies, pitch-synchronous frames are used, since this allows for prosodical manipulations as changing the fundamental frequency.

In this work, the impact of two variants of VC on the recognition accuracy was examined. In the first case, both fundamental frequency (F_0) and formants were transformed (converted LE speech, CLE), in the second only F_0 was converted and formants preserved (CLEf0). Distributions of F_0 and mean positions of the first two formants $F_{1,2}$ in the female digit vowels for neutral, LE, CLE and CLEf0 speech are shown in Fig. 3. The slight difference between LE and CLEf0 formant positions is presumably caused by occasional confusion of F_0 and F_1 by the automatic formant tracking algorithm.

2.4. Comparing performances

Firstly, MFCC features with modified FBs (LFCC, 20 bands and LFCC, 19 bands) and MR-ANN features were integrated into the HMM recognition system and compared to MFCC and PLP [14] features in the digit recognition task. Four neutral (1450 digits) and four LE (1880 digits) CLSD'05 sessions formed the open test set, see results in Tab. 1.

All newly proposed front ends outperformed the standard features on LE speech. The best results were reached by MR-ANN and LFCC, 20 band features, where performance on the LE set was significantly better and on neutral speech comparable to MFCC and PLP. It is shown in [5] that omitting low frequency components introduces a trade-off between performance on LE and neutral speech, which is also confirmed by the results in Tab. 1.

Conditions	Open set	
	Neutral	LE
MFCC	3.7	68.7
PLP	3.4	61.3
MR-ANN	4.1	42.1
LFCC, 20 bands, full band	3.3	49.4
LFCC, 19 bands, ≥ 625 Hz	6.6	24.6

Table 1: Comparing standard and new features in the digits task – word error rates – WER (%).

Secondly, the performance of the combination of voice conversion and feature extraction was evaluated. Expolog features [1] (designed to be more robust to LE than standard features), LFCC, 20 bands and RFCC features were compared to MFCC and PLP. The RFCC FB was tested in two feature extraction algorithms – as a replacement for the FB in MFCC (RFCC-DCTC) and for the FB in PLP (RFCC-LPC).

For this experiment, only utterances comprising sequences of eight digits were picked from the open test set for the evaluation, i.e. 768 digits in the neutral set and 1024 digits in each of the LE, CLE, and CLEf0 sets.

FE\Set	Neutral	LE	CLE	CLEf0
MFCC	3.7	71.3	30.9	58.6
PLP	2.9	47.4	25.2	49.0
Expolog	3.9	35.8	26.7	37.8
LFCC, 20 bands, full band	3.0	42.1	19.9	42.7
RFCC-DCTC	5.1	26.1	22.6	22.6
RFCC-LPC	4.6	23.0	23.1	23.7

Table 2: Combining voice conversion and feature extraction in the digit recognition task – WER (%).

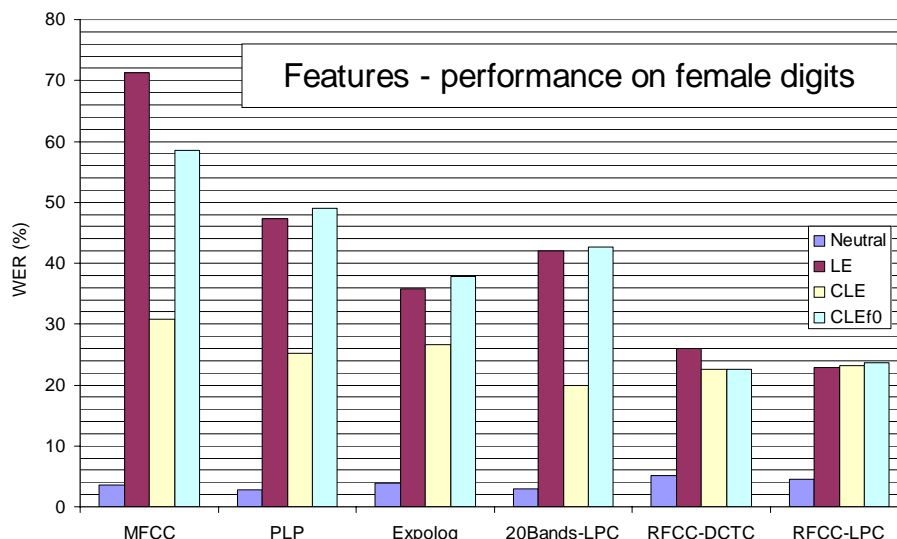


Figure 4: Combining voice conversion and feature extraction in the digit recognition task.

As shown in Tab. 2 and Fig. 4, with an exception of RFCC-LPC, VC improved recognition accuracy of all front ends (compar LE and CLE). Expolog features outperformed MFCC and PLP on the LE set while providing comparable accuracy on the neutral data. The best results on the LE set were achieved by RFCC-LPC. These features even gained on RFCC-DCTC, which lets us conclude that LPC is more suitable for loud speech modeling than DCT. This observation agrees with those reported in [1].

3. LVCSR task

The presented large LVCSR system [15] employs an MFCC front end. The acoustic part is formed by 48 HMMs, from which 41 represent Czech phonemes, the remaining 7 model noise events. All the models are context-independent and comprise a large number of Gaussian mixtures (up to 100). A multiple-pronunciation vocabulary with 312,764 words is used together with the corresponding bigram language model (LM). The LM is trained on the 4GB newspaper corpus [15].

Speech consisting of 973 words uttered by 11 male speakers and 970 words uttered by 11 female speakers was sampled with 16 kHz and recognized with and without LM using speaker-dependent acoustic models. As in the previous section, the impact of voice conversion on the recognition accuracy was evaluated, see Tab. 3 and Fig. 5.

Set	Neutral	LE	CLE	CLEf0
Male - no LM	77.9	85.3	91.5	85.5
Male - LM	40.4	55.8	69.4	60.0
Female - no LM	72.9	86.5	90.1	87.9
Female - LM	28.9	63.7	66.7	60.4

Table 3: Impact of LE and voice conversion on the LVCSR task – word error rates (%).

In all cases, the LM improved the performance. VC was not helpful in this task. Since in LVCSR a misrecognition of just one phoneme may result in a word mismatch, even a slight inaccuracy in the VC (see also Fig. 3) corrupts the recognition. As in the digits task, a significant decrease in performance for speech under LE was observed.

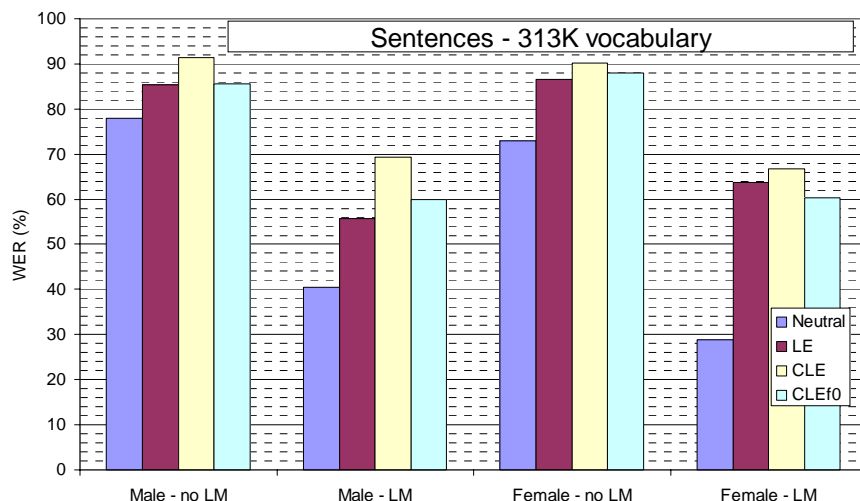


Figure 5: Impact of LE and voice conversion on the LVCSR task.

4. Model adaptation

For speaker adaptation, a combination of maximum a posteriori (MAP) [17] and maximum likelihood linear regression (MLLR) [18] methods was used. The adaptation was performed in two steps: Mean vectors of GMMs in speaker-independent (SI) acoustic HMMs were transformed by MLLR, and the transformed values were used as priors for the MAP-based adaptation. The implementation of MLLR used in this work was based on clustering of all acoustic models using a binary regression tree. During the adaptation process, the tree was searched down from the root towards the leaves while the transformations were calculated only for those nodes, where sufficient adaptation data was available. In the preliminary experiments, it was found that the clustering is more effective, if the first two nodes of the regression tree are created manually by splitting all the acoustic models into two groups: models of noises and models of phonemes. The benefit of this approach gains from the fact that models not covered in the adaptation data can still be adapted well by MLLR, while MAP only ensures that parameters of models with a lot of adaptation data can converge towards the theoretically best values. Only the mean vectors were adapted in this work. The following experiments were carried out on a set of 10 speakers:

- SI adapt to LE (same speakers train/test) – from all speakers, 2/3 of the LE utterances are used for SI model adaptation, 1/3 for open test (880 digits, 970 words)
- SI adapt to LE (different speakers train/test) – 6 speakers are used for SI model adaptation, the utterances from the remaining 4 speakers form the test set (1024 digits, 1081 words)
- SD adapt to neutral – speaker dependent (SD) models for each speaker are adapted to neutral utterances, tested on LE speech (880 digits, 970 words)
- SD adapt to LE – SD models are adapted to LE, 2/3 of LE utterances are used for SI model adaptation, 1/3 for open test (880 digits, 970 words)

The numbers in the above brackets represent the numbers of items being recognized using LM and a 312K vocabulary. The results of the experiments are in Tab. 4, 5 and Fig. 6.

Digits – WER (%)	LE baseline	LE adapted	Neutral baseline
SI adapt to LE (same speakers train/test)	54.7	16.8	15.0
SI adapt to LE (different speakers train/test)	55.5	16.9	–
SD adapt to neutral	54.7	43.9	15.0
SD adapt to LE	54.7	8.5	15.0

Table 4: Performance of acoustic models adaptation in the digit recognition – word error rates (%).

Sentences – WER (%)	LE baseline	LE adapted	Neutral baseline
SI adapt to LE (same speakers train/test)	69.7	43.0	32.3
SI adapt to LE (different speakers train/test)	78.5	61.2	–
SD adapt to neutral	69.7	68.7	32.3
SD adapt to LE	69.7	39.2	32.3

Table 5: Performance of acoustic models adaptation in the LVCSR task – word error rates (%).

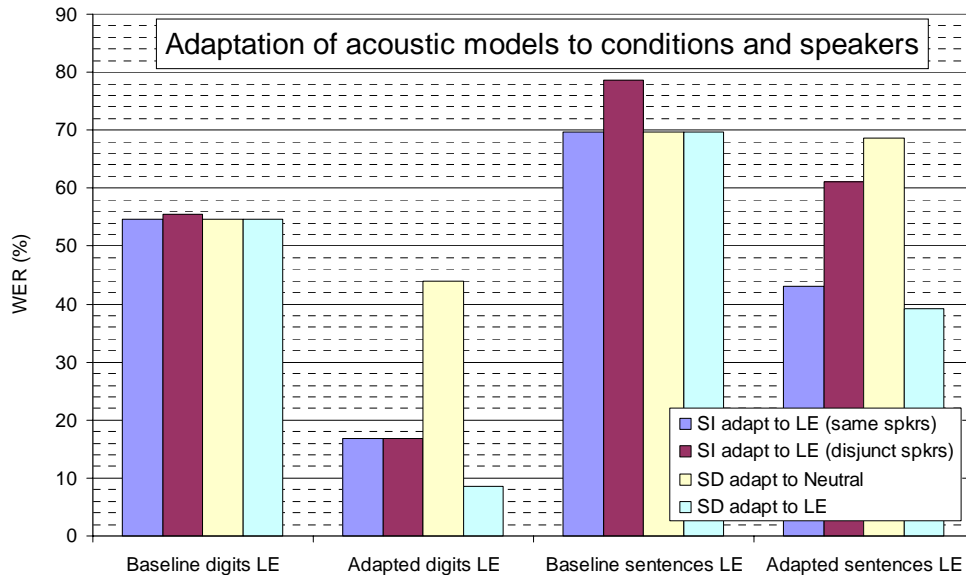


Figure 6: Efficiency of acoustic models adaptation in the digits and LVCSR task.

As shown in Tab. 4, 5 and Fig. 6, all configurations of the acoustic model adaptation improved the recognition performance. As expected, the best results were reached by SD model adaptation to LE. In general, various speakers react to the same noisy conditions differently, which may discourage any global adaptation. However, a remarkable improvement was observed for both SI adaptations to LE. Surprisingly, even the SI case with different speakers for the adaptation and evaluation improved the performance. SD adaptation to neutral speech achieved worse results than the other approaches, as the speech characteristics under LE differ significantly from the neutral ones (as shown also in Fig. 3).

5. Conclusions

Various approaches to Lombard speech recognition were introduced and evaluated. The performance of modified front-end filter banks, multi-resolution RASTA features, voice conversion, and acoustic model adaptation was tested on Czech SPEECON and CLSD'05 databases. Each of the mentioned approaches led to considerable improvements in performance. A trade-off was observed between tuning the system for the Lombard speech and the performance on neutral speech. It seems that small and large vocabulary tasks require different recognition approaches.

Acknowledgments

This work was supported by the grants GAČR 102/03/H085 “Biological and Speech Signals Modeling” and MSM 684077 0014 “Research in the Area of the Prospective Information and Navigation Technologies”. Robust FBs were designed within the joint project “Normalization

of Lombard Effect” of CTU Prague and Siemens Corporate Technology, initiated by Harald Höge. The acoustic model adaptation in the LVCSR task was supported by the grant GAČR 102/05/0278 “New Trends in Research and Application of Voice Technology”.

References

- [1] Bou-Ghazale, S. E., Hansen, J. H. L.: A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech under Stress. *IEEE Trans. on Speech and Audio Processing*, 8(4), 2000, pp. 429-442.
- [2] SPEECON Database, <http://www.speechdat.org/speecon>.
- [3] Bořil, H., Pollák, P.: Design and Collection of Czech Lombard Speech Database. *Proc. of Interspeech'05*, Lisbon, Portugal, 2005, pp. 1577-1580.
- [4] Mermelstein, P., Davis, S.: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. on Acoustic, Speech, and Signal Processing*, 28(4), 1980, pp. 357–366.
- [5] Bořil, H., Fousek, P., Pollák, P.: Data-Driven Design of Front-End Filter Bank for Lombard Speech Recognition. *Proc. Interspeech'06*, Pittsburgh, 2006, pp. 381-384.
- [6] Hermansky, H., Fousek, P.: Multi-Resolution RASTA Filtering for TANDEM-Based ASR. *Proc. of Interspeech'05*, Lisbon, Portugal, 2005, pp. 361-364.
- [7] Sündermann, D.: Voice Conversion: State-of-the-Art and Future Work. *Proc. of DAGA'05*, Munich, Germany, 2005.
- [8] Sündermann, D., Höge, H., Bonafonte, A., Ney, H., Hirschberg, J.: Text-Independent Cross-Language Voice Conversion. *Proc. of Interspeech'06*, Pittsburgh, USA, 2006.
- [9] Goncharoff, V., Gries, P.: An Algorithm for Accurately Marking Pitch Pulses in Speech Signals. *Proc. of SIP'98*, Las Vegas, USA, 1998.
- [10] Paliwal, K.: Interpolation Properties of Linear Prediction Parametric Representations. *Proc. of Eurospeech'95*, Madrid, Spain, 1995.
- [11] Kain, A., Macon, M.: Spectral Voice Conversion for Text-to-Speech Synthesis. *Proc. of ICASSP'98*, Seattle, USA, 1998.
- [12] Sündermann, D., Bonafonte, A., Ney, H., Höge, H.: A Study on Residual Prediction Techniques for Voice Conversion. *Proc. of ICASSP'05*, Philadelphia, USA, 2005.
- [13] Charpentier, F., Stella, M.: Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation. *Proc. of ICASSP'86*, Tokyo, Japan, 1986.
- [14] Hermansky, H.: Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal of the Acoustic Society of America*, 87(4), 1990, pp. 1738-1752.
- [15] Nouza, J., Zdansky, J., Cerva, P., Kolorenc, J.: Continual On-line Monitoring of Czech Spoken Broadcast Programs. *Proc. of Interspeech'06*, Pittsburgh, USA, 2006.
- [17] Gauvain, J. L., Lee, C. H.: Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Trans. on Speech and Audio Processing*, 2, 1994, pp. 291-298.
- [18] Gales, M. J. F., Woodland P. C.: Mean and Variance Adaptation within the MLLR Framework. *Computer Speech & Language*, 10, 1996, pp. 249-264.