# Data-Driven Design of Front-End Filter Bank for Lombard Speech Recognition

*Hynek Bořil, Petr Fousek, Petr Pollák*

Czech Technical University in Prague, Faculty of Electrical Engineering,
Prague, Czech Republic

{borilh,p.fousek}@gmail.com, pollak@feld.cvut.cz

## Abstract

Adverse environments not only corrupt speech signal by additive and convolutional noises, which can be successfully addressed by a number of suppression algorithms, but also affect the way how speech is produced. Speech production variations introduced by a speaker in reaction to a noisy background (Lombard effect) may result in a severe degradation of automatic speech recognition. This paper contributes to the solution of Lombard speech recognition issue by providing a robust filter bank for use in front-ends. It is shown that cepstral features derived from the proposed filter bank significantly outperform conventional cepstral features.

**Index Terms**: robust speech recognition, Lombard effect, feature extraction, filter bank, data-driven design

## 1. Introduction

Thanks to a huge effort invested in development of speech recognition systems, current ASR deals well with a speech produced in a quiet environment. Also a vast number of algorithms targeting adverse environments have been developed recently, successfully addressing additive and convolutional noise. In noisy conditions, humans tend to modify their speech production in an effort to preserve an intelligibility of communication (Lombard effect, LE) [1]. LE introduces significant deviations in distribution of speech parameters crucial for automatic recognition, e.g. shifts of fundamental frequency ($f_0$), formant frequencies and bandwidths, resulting in severe degradation in performance of a recognizer trained on neutral speech. Up to now, several approaches solving particular topics in LE have been proposed, yet a complex solution remains unattained.

LE suppression can be addressed from three perspectives: robust feature extraction, LE equalization, or model adjustment by multi-style training. The multi-style training performs well in speaker and environment specific task. However, for varying speakers and environments the multi-style trained recognizer fails, since the parameter changes are strongly dependent on the particular conditions and can hardly be represented completely in a limited training data [1]. This work focuses on the robust feature extraction approach.

Majority of current ASR systems employ MFCC [2] or PLP [3] features for their superior properties to previously used representations (LPCC). One of the key processing stages common to both algorithms is a smoothing of FFT spectrum with a bank of nonlinearly distributed filters. Their distribution is derived from auditory models in an effort to emphasize the similar speech components that are essential for human speech perception. Some works have reached an improvement by further modifying auditory based filter banks (FBs), e.g. Human Factor Cepstral Coef-

| Conditions | Devel Set | | Open Set | |
|---|---|---|---|---|
| | Neutral | LE | Neutral | LE |
| MFCC | 3.6 | 63.3 | 3.7 | 68.7 |
| PLP | 3.8 | 54.1 | 3.4 | 61.3 |
| MR-ANN | 4.5 | 39.8 | 4.1 | 42.1 |

Tab. 1: Word error rates (%) of baseline features on female neutral and Lombard speech, development and open test sets.

ficients (HFCC) changing bandwidths of mel filters [4]. Others proposed new auditory models, e.g. Seneff auditory model comprising 40 filters matching cat's basilar membrane response [5] or Ensemble Interval Histogram (EIH) model employing a bank of level crossing intervals [6]. Also non-auditory, data-driven concepts of FB design were studied, e.g. Discriminative Feature Extraction method (DFE), iteratively adapting FB parameters [7] or a design of a library of phoneme class-dependent filter banks using F-ratio [8]. Filter banks introduced in these works were tested in simulated noisy conditions, yet no extensive research on Lombard effect has been reported. [9] tested various FBs with a simulated loud utterances, though not all properties of real loud speech have been considered (e.g. $f_0$ and formant shifts).

Suitability of various features for recognition of speech covering different talking styles was studied in [1]. Further, the mel FB was adjusted to enhance stressed speech recognition (including LE). Inspired by these experiments, a goal of this work is to design a novel filter bank which would improve feature robustness and ASR performance in presence of LE. As comparative experiments show significantly stronger corruption of recognition in case of female than male Lombard speech [10] (see also Sec. 3), all experiments presented in this work were carried out for female speech only.

The paper is organized as follows. First, common features are compared on a digits recognition task. In addition, a recently proposed Multi-resolution RASTA features [11] participate in the tests. Second, an importance of frequency subbands for recognition is explored and further used in the process of designing a new FB. Third, a relation between the importance of spectral components and corresponding FB resolution is examined. Finally, an iterative algorithm of repartitioning bands in FB is proposed and evaluated.

## 2. Development setup

### 2.1. Used corpora

All experiments in this paper were carried out on Czech SPEECON [12] and CLSD'05 [13] corpora. Czech SPEECON

September 17–21, Pittsburgh, Pennsylvania

comprises recordings in public, office, car and entertainment scenarios. CLSD'05 contains speech uttered in various types of simulated noisy background (car and artificial band-noises). Significant $f_0$ and format frequencies shifts were observed in the database for simulated LE conditions [13]. Since the noise was reproduced to speakers through closed headphones, a clean Lombard speech was captured at high SNR.

For HMM training, office data from SPEECON with neutral speech in a quiet environment were used. This set contained general speech pronounced by females and covered full phonetic content of Czech language. For the development and open testing, disjunct sets from SPEECON and CLSD'05 data were used, consisting of female neutral and LE speech, respectively. All data were downsampled to 8 kHz and filtered by G.712 telephone filter using FaNT tool [14]. Overview of train and test sets:

**Train** – 10 hours of signal, 37 female speakers

**Devel LE** – 3480 words, 8 female speakers,

**Devel neutral** – 3480 words, 8 female speakers,

**Open LE** – 1880 words, 4 female speakers,

**Open neutral** – 1450 words, 4 female speakers.

The recognizer was an HMM system with 43 context-independent phoneme models + 2 silences, each with 3 emitting states and 32 Gaussian mixtures per state. The task was to recognize 10 Czech digits in 16 pronunciation variants.

## 3. Baseline front-ends performance on LE speech

To get an idea about a performance of various speech representations on the given corpora, two common feature sets plus a new posterior-based features obtained from neural network were compared:

- MFCC – 3×13 coeffs, 26 bands, 100 Hz frame rate, preemphasis, liftering,

- PLP – 3×13 coeffs, 26 bands, mel scale, 100 Hz frame rate, preemphasis, liftering,

- MR-ANN – 39 coeffs derived from 43 phoneme posteriors.

Recently proposed Multi-resolution RASTA features (MR-ANN) [11] are derived in two steps. In the first step, energies of the speech signal in auditory subbands are computed. One second long trajectories of these energies surrounding the point of interest form an auditory spectrogram. This spectrogram is further filtered with a bank of two-dimensional filters, yielding a set of about 500 numbers at each speech frame. In the second step an artificial neural network is used to estimate posterior probabilities of phonemes given the set of 500 numbers, reducing the feature size. These posteriors are finally decorrelated using principal components analysis to be able to fit the GMM/HMM model.

All features comprising PLP and MFCC baselines and subband energies for MR-ANN were extracted using an open source speech enhancement and parametrization tool CtuCopy developed at CTU in Prague [15]. CtuCopy also implemented the newly proposed filter banks.

Performances of the above mentioned systems are summarized in Tab. 1. Clean sets establish a baseline at about 4% WER.

On LE data a huge decrease in accuracy was observed for all features. MFCC displayed the worst results, MR-ANN significantly outperformed both MFCC and PLP.

For a comparison, similar training and testing sets were designed for a male speech. On this gender the recognizer performed much better in case of Lombard speech: 29.7 % WER for MFCC, 25.4 % WER for PLP.
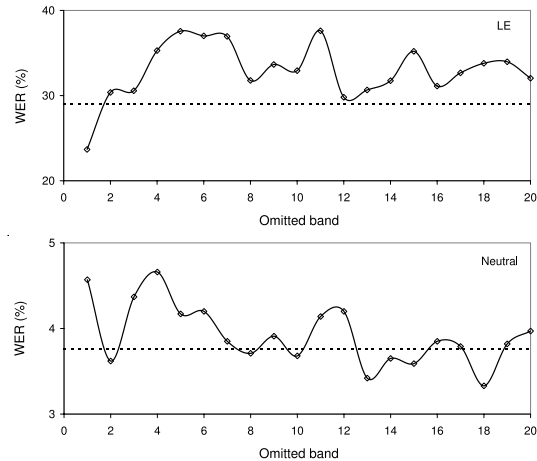


Fig. 1: ID curves: impact of one missing band in 20–band FB on recognition performance (devel set).

## 4. Designing filter bank

### 4.1. Analyzing importance of subbands

Knowledge about spectral distribution of linguistic message in the speech signal may provide a useful guideline for a FB construction. In [1], a set of recognizers was trained and tested using individual band outputs, hence the spectral envelope and dependencies between bands (formant distances and bandwidths) were not considered. A new FB was then constructed given the distribution of scores in independent bands and an assumption that the filter density should be raised in regions carrying more information.

The work presented here estimated a score-based information distribution across frequency (ID) by keeping all filters in FB but the examined one. In general, omitting one band can either rise the score compared to the baseline (the baseline is depicted by dashed line in all figures), meaning that the band is dominated by irrelevant information, or decrease the score proving the band's importance for the recognition.

The initial FB was chosen to consist of 20 linearly spaced rectangular filters (each of bandwidth 200 Hz) without overlap. For LE speech, the baseline score on the development set reached word error rate (WER) 29.0 %, which significantly outperformed features presented in the previous section. As shown in Fig. 1, omitting the first band brings a slight degradation on neutral speech, but greatly enhances LE recognition (see also the first row in Tab. 2).

In case of LE speech, a significant peak in the ID curve can be observed in the region 600 - 1400 Hz (bands 4 - 7), covering an area of occurence of the first two formants. For neutral speech, a corresponding peak lies in 400 - 1000 Hz (bands 3 - 5), the area of the first formant location. This agrees with the conclusions drawn for angry speech by [1], where the highest recognition per-

formance for neutral speech was observed around the first formant location, while for angry speech the maximum moved rather to the area of the second formant. Fig. 1 also suggests that Lombard speech recognition may be improved by avoiding low-frequency components at the expense of neutral speech recognition accuracy.

A similar experiment was carried out on 43 bands FB but the ID was noisy as the omitted bands were probably too narrow to noticeably affect the information content.

### 4.2. Avoiding low-frequency components

Previous section mentioned an improvement trade-off between neutral and Lombard speech when avoiding low-frequency components. As a number of efficient features are available for neutral speech, following design steps focused exclusively on LE recognition improvement. Hence, a dependency between the low cut-off frequency and recognition score was explored, see Fig. 2. The minimum of WER on Lombard speech was found at 625 Hz, rising the score by 13.4 %. The decrease on neutral speech (1.8 %) was proportional to the cut-off frequency.
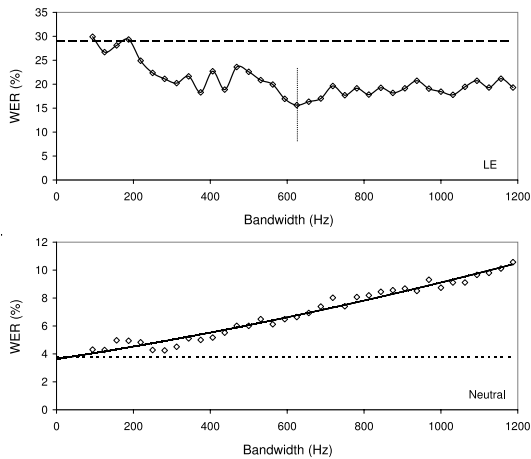


Fig. 2: Searching for optimal low cut-off frequency in 19–band FB: increasing significantly improves Lombard speech recognition, performance on neutral speech decreases (devel set).

|            | Devel Set |      |
| ---------- | --------- | ---- |
| Conditions | Neutral   | LE   |
| LFCC, Full Band | 4.8  | 29.0 |
| LFCC, $\geq$ 625 Hz | 6.6 | 15.6 |

Tab. 2: Word error rates (%) of cepstra derived from a bank of linearly spaced rectangular filters (LFCC): (1) 20 filters, 0–4000 Hz, (2) 19 filters, 625–4000 Hz.

### 4.3. Importance and resolution

To obtain a smoothed ID curve for the FB starting at the optimized cut-off frequency, the number of bands was lowered to 12. Fig. 3 shows that a higher importance is assigned to low FB bands, which is in agreement with the reported relevance of the first two formants to the speech recognition [1]. In the following step, the first

band in FB was split in two and two subsequent bands were replaced with three bands in order to increase the resolution at low frequencies. By this modification, the score dropped from 17.2 % to 26.9 % WER for LE. It suprisingly suggests that increasing the resolution in bands of higher importance does not necessarily improve the system.

|            | Devel Set |      |
| ---------- | --------- | ---- |
| Conditions | Neutral   | LE   |
| LFCC, 19 bands | 6.6   | 15.6 |
| LFCC, 12 bands | 8.0   | 17.2 |
| LFCC, 6 bands  | 9.6   | 17.9 |

Tab. 3: Word error rates (%) of cepstra derived from a bank of linearly spaced rectangular filters (LFCC), different resolutions.

### 4.4. Filter bank repartitioning algorithm

The experiment with changing the FB resolution at low frequencies demonstrated that it is not possible to design a LE-robust FB just by modifying the distribution of bands in FB according to the ID curve. At the same time, it has been shown that filter bandwidths significantly impact accuracy. This observation motivated a development of a FB repartitioning algorithm.

The idea is to search for an optimal bandwidth of each filter while leaving the rest of FB intact as much as possible. In the first step, the endpoint of the first filter varied around its original location and a new position yielding minimal WER was searched iteratively. For each endpoint position the rest of FB was resized to preserve equal bandwidths of the remaining filters. Once the optimum for the band was reached, it was fixed. In the subsequent steps, successive bands were processed the same way, keeping the previously fixed points intact.

For testing purposes, number of FB bands was limited to 6 in order to reduce the computational costs. FB reduction inherently affected baseline performance, see Tab. 3. Nevertheless, the repartitioned FB improved on the baseline by 2.3 % for LE speech, as shown in Tab. 4. The resulting FB started at 625 Hz and the endpoints of individual bands were 1125, 1719, 2313, 2875, 3438 and 4000 Hz.

## 5. Summary

Widely used MFCC and PLP speech features were exposed to a strong presence of Lombard effect in a digit recognition task,
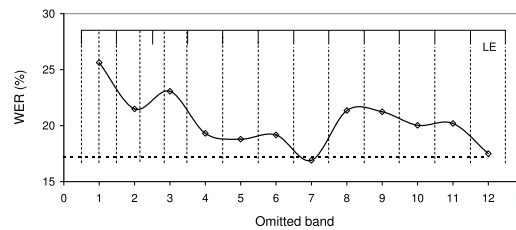


Fig. 3: Increasing FB resolution in region displaying superior importance. Solid lines denote the former 12-band FB, dashed lines the resulting 14–bands FB (devel set).
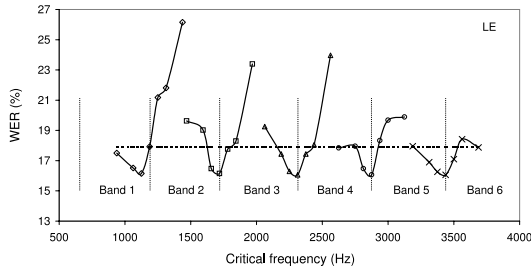
Fig. 4: Search of optimal band partitioning for 6–band FB. For each band sequentially, endpoint yielding best performance is found, preserving distribution of preceding bands (devel set).

| 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, $\geq$625 Hz | 6.6 | 24.6 |
| LFCC, 12 bands, $\geq$625 Hz | 7.4 | 25.6 |
| LFCC, 6 bands, $\geq$625 Hz | 9.5 | 31.7 |
| RFCC, 6 bands, $\geq$625 Hz | 8.5 | 29.4 |

Tab. 4: Evaluation of all systems on independent open test set, word error rates (%). Compared systems were: MFCC, PLP, Multi-RASTA neural network (MR-ANN), cepstra from linearly spaced rectangular filters (LFCC) and repartitioned filters (RFCC).

which revealed their poor recognition performance. Recently proposed features based on an artificial neural network (Multi-resolution RASTA) achieved substantionaly better performance on LE speech (about 20 %), though still not satisfactory for real life aplications. Among possible approaches addressing Lombard speech recognition, this work investigates on robust feature extraction, particularly a filter bank design. The priority is to improve on Lombard speech recognition.

The prototype filter bank was formed by non-overlapping rectangular filters of equal bandwidths mapped to a linear frequency scale to ensure initial equality of their contributions. Preliminary evaluations showed superior performance of the prototype filter bank based cepstral features when compared to MFCC and PLP, considering both speech conditions (about 15 % on Lombard speech).

Prior to modifying the filter bank, an independent contribution of spectral components of speech to the recognition was evaluated. Results suggested omitting the low frequency components which further improved the accuracy by 25 %. It was experimentally shown that increasing the resolution of filter bank in regions of a higher importance does not necessarily improve the system, as the score dropped by 10 %.

An observation that filter bandwidths impact accuracy significantly was a motivation for developing a filter bank repartitioning algorithm. The proposed algorithm was evaluated using a simplified filter bank, yielding an additional improvement of 2 %.

## 6. Acknowledgements

## 7. References

[1] Bou-Ghazale, S. E., J.H.L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," IEEE Trans. on Speech & Audio Processing, vol. 8, no. 4, pp. 429-442, July 2000.

[2] Mermelstein, P., S. Davis, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. on Acoustic Speech and Signal Processing, 28(4), pp. 357–366, August 1980.

[3] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech," JASA, Vol. 87, No. 4, April 1990, p. 1738-1752.

[4] Skowronski, M. D., J. G. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," JASA, vol. 116, no. 3, pp. 1774-1780, Sept. 2004.

[5] Seneff, S., "A computational model for the peripheral auditory system: Application to speech recognition research," Proc. of ICSLP'86, Tokyo, pp. 1983–1986.

[6] Ghitza, O., "Auditory nerve representation as a basis for speech processing," Advances in Speech Signal Processing, N.Y., 1992, pp. 453–486.

[7] Biem, A., S. Katagiri, "Cepstrum-based filter-bank design using discriminative feature extraction training at various levels," Proc. of ICASSP'97, p. 1503, Volume 2, 1997.

[8] Kinnunen, T., "Designing a speaker-discriminative adaptive filter bank for speaker recognition", Proc. of ICSLP'02, pp. 2325-2328, Denver, Colorado, USA, 2002.

[9] Jankowski Jr., C. R., et al. "A Comparison of signal processing front ends for automatic word recognition," IEEE Trans. on Speech an Audio Processing, Vol. 3(4), July 1995, pp. 286–293.

[10] Junqua, J.C., "The Lombard reflex and its role on human listeners and automatic speech recognizers," JASA, 93(1):637–642, 1993.

[11] Hermansky, H., P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," Proc. of Interspeech'05, Lisbon, Portugal, 2005.

[12] SPEECON database, <http://www.speechdat.org/speecon>.

[13] Bořil, H., P. Pollák , "Design and collection of Czech Lombard Speech Database," Proc. of Interspeech'05, Lisboa, Portugal, 2005, p. 1577 - 1580.

[14] FaNT - Filtering and Noise Adding Tool. <http://dnt.kr.hsnr.de/download.html>.

[15] Fousek, P., P. Pollák, "Additive noise and channel distortion-robust parametrization tool - performance evaluation on aurora 2 & 3," Proc. of Interspeech'03, Geneve, Switzerland, 2003.