Influence of Different Speech Representations and HMM Training Strategies on ASR Performance

H. Bořil, P. Fousek

This work studies the influence of various speech signal representations and speaking styles on the performance of automatic speech recognition (ASR). The efficiency of two approaches to hidden Markov model (HMM) training are compared.
Common MFCC and PLP features were exposed to two sources of disturbance applied to the original wide-band speech: (i) stress (Lombard effect) and (ii) transfer channel distortion (simulated telephone line). Subsequently, the efficiencies of the two training strategies were evaluated. Finally, a study of the optimal number of training iterations is introduced.

Keywords: PLP, MFCC, Lombard effect, CLSD'05.

This text was a part of the International Conference POSTER 2006 which was held in Faculty of Electrical Engineering CTU in Prague.

1 Introduction

The recognition of clean speech recorded in quiet conditions can be addressed quite successfully with the widely used Mel-Frequency Cepstral Coefficients – MFCC [1] and Perceptual Linear Predictive Coefficients – PLP [2]. In this work, the behavior of these features is examined in the case of changes in talking style (neutral speech, speech under Lombard effect – LE) and in the case of speech bandwidth limitation introduced by telephone filter. LE introduces changes in speech production due to the speaker's effort to increase communication intelligibility in noise [3]. Bandwidth limitation of telephone filter introduces changes in the spectral content of the signal which may lead to loss of the fourth speech formant and may thus degrade the recognizer performance.

In addition, two HMM training strategies were compared with respect to convergence speed and best achievable performance. The strategies differed only in the way in which the initial HMMs containing one Gaussian mixture component per HMM state were enhanced to contain the final 32 mixtures per HMM state. This was done either by direct splitting and cloning the only mixture to 32 mixtures and reestimating them many times (*one-shot* approach), or by gradually doubling the number of mixtures and reestimating after each split until there were 32 mixtures (*progressive propagation*).

The recognition experiments presented in this paper were carried out on Czech SPEECON [4] and CLSD'05 [5] databases. Czech SPEECON comprises recordings in public, office, car and entertainment scenarios. For the purpose of HMM training, office data representing neutral speech with high SNR was chosen. For the tests, the CLSD'05 database was used. This consists of neutral speech and speech uttered in various types of simulated noisy backgrounds (CAR2E car noise [6] artificial band-noises). Since the noises were reproduced to the speakers through closed headphones, only clean Lombard speech was captured, benefiting from similar SNR to the SPEECON office recordings.

2 Feature extraction techniques

Two widely used feature extraction methods were examined in this work, MFCC and PLP. Both methods use Mel-Frequency warping with the same number of frequency subbands, and the estimates of the spectral envelopes were described by the same number of cepstral coefficients, so that the two methods were comparable. The most important difference in the otherwise rather similar approaches is the way in which the cepstral coefficients are obtained from the spectra: through the DCT transform in the case of MFCC, and through linear prediction in the case of PLP. The settings were as follows:

- preemphasis with $\alpha = 0.97$
- 100 Hz frame rate, frame length 25 ms
- 26 Mel-frequency bands
- LPC of order 12 (for PLP)
- energy normalization per utterance
- cepstral liftering
- 39 features per frame (12 cepstral coeffs + frame energy, Δ and $\Delta\Delta$ coeffs)

The input speech data was either sampled at 16 kHz or resampled at 8 kHz. As the feature extraction settings were not dependent on the sampling rate, the frequency subbands were effectively wider for 16 kHz than for 8 kHz.

3 Telephone channel simulation

One of the main goals of the study was to investigate the influence of limiting the bandwidth of the input speech, particularly by simulating a standard telephone channel. The

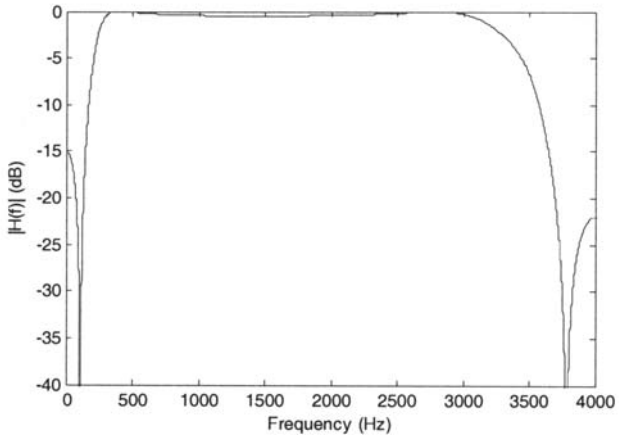

Fig. 1.: Transfer function of the simulated telephone channel

 © Czech Technical University Publishing House

original 16 kHz, 16 bit PCM speech was processed in two steps. First, the signal was resampled using the sox tool with a polyphase filter to 8 kHz [7]. Then the telephone channel was emulated by applying a G.712 standard IIR filter of the 4th order [8] using the FaNT tool [9]. Superposition of the processing steps as described above leads to an effective bandwidth of 300 Hz–3400 Hz (see Fig. 1).

4 Recognition experiments

4.1 Experimental setup

All experiments were carried out on the Czech SPEECON and CLSD'05 databases. From both databases only the close-talk microphone channel was used. The training set consisted of SPEECON office recordings, which represent neutral speech in a quiet environment. This set contained general speech pronounced by both genders, and comprised about 15 hours of speech. There were four independent test sets covering examples of gender-dependent neutral or Lombard speech: neutral-male (1423 words), neutral-female (4930 words), LE-male (6303 words) and LE-female (5360 words), containing continuously pronounced digits from "nula" to "devět". Though the neutral and Lombard utterances differed in the prompt texts, speakers were the same for both sets.

The recognizer was a gender-independent HTK-based 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.

4.2 Effect of Lombard speech and resampling on MFCC and PLP features

The aim of this experiment was to show how the Lombard effect and a narrow bandwidth can affect recognition performance. In all cases, the training and testing conditions were the same.

First, the baseline performance of MFCC and PLP features on neutral wide-band speech was evaluated in terms of Word Error rate (WER), see rows 1–2, columns 1–2 in Table 1. Also, similar narrow-band systems were tested (rows 3–4, columns 1–2 of Table 1). Then all four systems were exposed to Lombard speech (columns 3–4 of Table 1). The observations are:

- MFCC and PLP features display comparable performance in all conditions.
- The Lombard effect leads to severe but consistent degradation: the relative drop from neutral to Lombard speech is comparable for male and female (about 800 %) and almost independent of bandwidth and features. However, the absolute errors indicate that for female speakers the recognizer is almost useless.
- Narrowing the bandwidth to the telephone introduces a degradation which is consistent over gender, features and speaking style. On average there is a relative drop of around 30 %. Note the special case of PLP features and neutral female speech, when the relative drop is only 12 %.

To help in interpreting the above observations, two phenomena should be mentioned. First, a known property of the

Table 1: Gender-dependent recognition results with neutral and Lombard speech at different bandwidths

Features	bandwidth	Word Error Rate (%)			
		Neutral		Lombard	
		Male	Female	Male	Female
MFCC	wide	2.4	4.9	18.8	43.9
PLP	wide	2.5	5.0	18.6	44.9
MFCC	telephone	3.0	6.2	24.2	62.2
PLP	telephone	3.2	5.6	23.1	62.5

Lombard effect is a significant shift of the first two formant frequencies [10], see Fig. 2. This may cause inability of the Gaussian mixtures to match the testing data and thus failure of the system. Avoiding this can be helped by appropriate front-end processing (equalization of LE, robust features), multi-style training (including Lombard speech in the training data) or back-end processing (changes of HMM structure) [3].

Second, the formants carry important information for monophone identification [11]. Narrowing the bandwidth to the telephone channel causes a loss of the 4th formant, which is close to 4 kHz (see Table 2). This can contribute to a performance drop in narrow band systems.

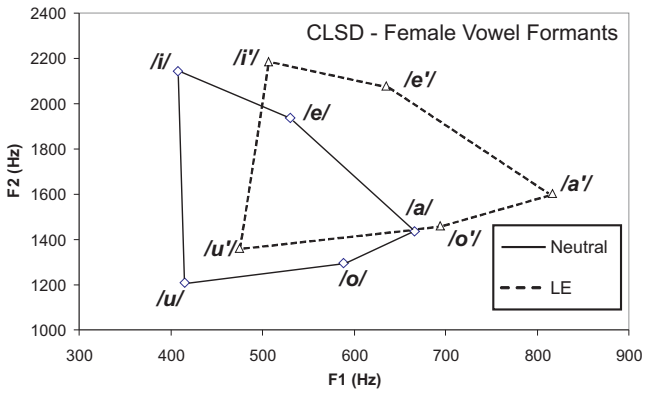

Fig. 2: CLSD'05 – vowel formant shifts

Table 2: CLSD'05 – average positions of 4th formants in neutral and Lombard speech

Vowel	Neutral		Lombard	
	F_{4Male} (Hz)	$F_{4Female}$ (Hz)	F_{4Male} (Hz)	$F_{4Female}$ (Hz)
/a/	3834	3934	3713	4012
/e/	3696	4181	3728	4196
/i/	3661	4170	3683	4218
/o/	3916	3880	3711	4042
/u/	3738	3939	3661	4001

4.3 Comparing training strategies

All the recognizers mentioned up to now were trained using the progressive propagation method: initial HMMs containing one Gaussian mixture (GM) per state were re-estimated using the Baum-Welch procedure and then each GM was split into two GMs and reestimated. After 5 cycles there were 32 GMs, which were further trained. This experiment compares such an approach with the one-shot strategy, where the initial mixture was cloned 32 times in each HMM to create 32 GMs directly. The HMMs were then reestimated.

The performance of both strategies was tested on a set comprising 8279 digits from the SPEECON office and CLSD'05 neutral sessions, see Fig. 3.

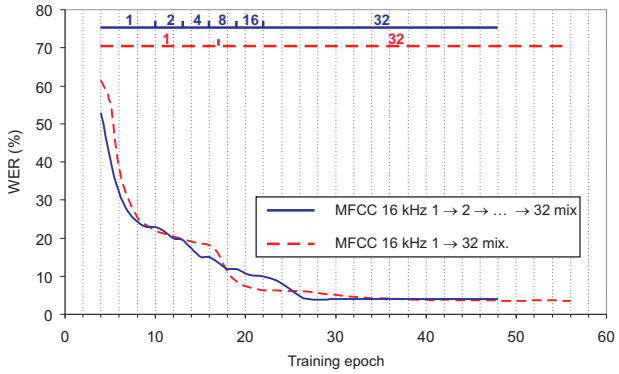

Fig. 3: Comparing training strategies – progressive propagation (blue) vs. one-shot (red). The top axes show the number of mixtures in an epoch.

To complete the picture about the training process, the evolution of insertions and deletions is shown in Fig. 4. No word insertion penalty was used.

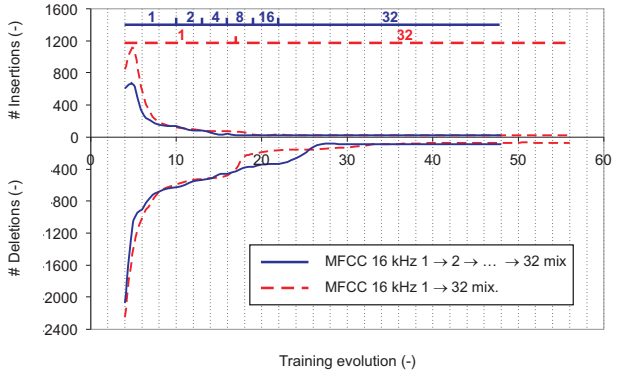

Fig. 4: Comparing training strategies – convergence of word insertions and deletions

4.4 When to stop training

The last experiment attempts to answer the following questions: How many training iterations should be performed in order to get the best models? Are *the best* models for neutral speech also *the best* for Lombard speech? The wide-band MFCC system trained with *progressive propagation* was used to recognize neutral and Lombard speech in each training epoch. Fig. 5 shows the performance evolution.

HMMs tested with neutral speech appear to converge much earlier than with Lombard speech. Excessive reestimations improve the performance on Lombard speech and do not seem to harm neutral speech. This suggests that many

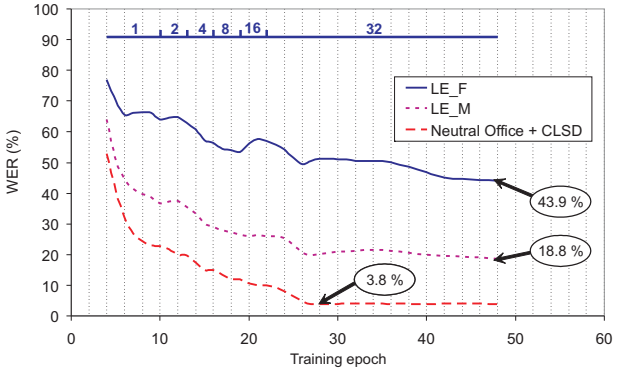

Fig. 5: Progressive propagation training – evolution of WER on male/female Lombard speech and both-genders neutral speech

iterations do not lead to loss of the essential generalization properties of HMMs.

5 Conclusions

The aim of the paper was to study the effect of narrowing the speech bandwidth and the effect of stressed speech on the performance of the HMM recognizer based on MFCC and PLP features. Experiments were carried out with the Czech SPEECON and CLSD'05 corpora.

MFCC and PLP features displayed similar behavior in all conditions. No fundamental differences were observed.

Narrowing the bandwidth to the telephone channel brought performance deterioration, which was consistent over gender, features and speaking style. A possible explanation is the loss of the 4th speech formant.

A consequence of the Lombard effect was a severe drop in performance, common to both features. Though the relative drop was comparable for both genders and bandwidths, in the female case it led to a failure of the recognizer. Without appropriate modifications, an HMM recognizer is almost useless when exposed to Lombard speech.

A comparison of the two training strategies showed their similar behavior and thus there is no need for further exploration.

An experiment with a higher number of HMM training iterations indicated that in order to achieve better recognition ccuracy on stressed speech, more training epochs are needed. Fortunately, these iterations do not damage the necessary generalization properties of HMMs.

Acknowledgments

This work was supported by GAČR 102/05/0278 "New Trends in Research and Application of Voice Technology", GAČR 102/03/H085 "Biological and Speech Signals Modeling", and research activity MSM 6840770014 "Research in the Area of Prospective Information and Navigation Technologies".

 © Czech Technical University Publishing House

References

[1] Young, S. et al.: *The HTK Book ver. 2.2*. Entropic Ltd 1999.

[2] Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech, *J. Acoust. Soc. Am*., Vol. **87**, No. 4, April 1990, p. 1738–1752.

[3] Hansen, J. H. L.: Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communications, Special Issue on Speech under Stress*, Vol. **20** (1996), No.2, p. 151–170.

[4] SPEECON, http://www.speechdat.org/speecon.

[5] Bořil, H., Pollák, P.: Design and collection of Czech Lombard Speech Database. In: *Proc. INTERSPEECH '05*. Lisboa (Portugal), 2005, p. 1577–1580.

[6] Pollák, P., Vopička, J., Sovka, P.: Czech language database of car speech and environmental noise. In *Proc. EUROSPEECH '99*. Budapest (Hungary) 1999. Vol. 5, p. 2263–6.

[7] SOX – Sound Exchange Tool manual, http://sox.sourceforge.net.

[8] The International Telegraph and Telephone Consultative Committee (CCITT), International Telecommunication Union (ITU). CCITT G.712: General Aspects of Digital Transmission Systems; Terminal Equipments. *Transmission Performance Characteristics of Pulse Code Modulation*, 1992.

[9] FaNT – Filtering and Noise Adding Tool. http://dnt.kr.hsnr.de/download.html.

[10] Bořil, H., Pollák, P.: Comparison of three Czech speech databases from the standpoint of Lombard effect appearance. In: *ASIDE 2005 – Applied Spoken Language Interaction in Distributed Environments*. Aalborg (Denmark), 2005. International Speech Communication Association. Book of abstracts [CD-ROM].

[11] Rabiner, L.R., Schafer, R. W.: *Digital Processing of Speech Signals*. Prentice Hall, New Jersey, 1978.

Ing. Hynek Bořil
e-mail: borilh@gmail.com

Petr Fousek
e-mail: p.fousek@gmail.com

Department of Circuit Theory

Czech Technical University
Faculty of Electrical Engineering
Technická 2
166 27 Prague, Czech Republic

© Czech Technical University Publishing House