Two-Stage System for Robust Neutral/Lombard Speech Recognition

Hynek Bořil [1], *Petr Fousek* [1], *Harald Höge* [2]

[1] Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic
[2] Siemens Corporate Technology, Munich, Germany

{borilh, p.fousek}@gmail.com, harald.hoege@siemens.com

Abstract

Performance of current speech recognition systems is significantly deteriorated when exposed to strongly noisy environment. It can be attributed to background noise and Lombard effect (LE). Attempts for LE-robust systems often display a tradeoff between LE-specific improvements and the portability to neutral speech. Therefore, towards LE-robust recognition, it seems effective to use a set of conditions-dedicated subsystems driven by a condition classifier, rather than attempting for one universal recognizer.

Presented paper focuses on a design of a two-stage recognition system (TSR) comprising talking style classifier (neutral/LE) followed by two style-dedicated recognizers differing in input features. First, the binary neutral/LE classifier is built, with a particular interest in developing suitable features for the classification. Second, performance of common speech features (MFCC, PLP), LE-robust features (Expolog) and newly proposed features is compared in neutral/LE digit recognition tasks. In addition, robustness to the changes of average speech pitch and various noise backgrounds is evaluated. Third, the TSR is built, employing two recognizers, each using style-specific features. Comparison of the proposed system with either neutral-specific or LE-specific recognizer on a joint neutral/LE speech shows an improvement 6.5→4.2 % WER on neutral and 48.1→28.4 % WER on LE Czech utterances.

Index Terms: Lombard effect, talking style classification, robust features, speech recognition

1. Introduction

Lombard effect refers to changes in speech production introduced by speaker in an effort to maintain intelligible communication [1, 2]. Number of works has studied impact of noise on speech production. Some analyzed acoustic-phonetic variations in few discrete levels of noise background [1, 2, 4, 9, 10], others searched for a continuous dependency on the noise level [11, 12]. Significant differences in distributions of vocal intensity, fundamental frequency, glottal pulse shape and spectral tilt, locations and bandwidths of first formants, and other parameters were reported between LE and neutral speech [1], substantially impairing accuracy of recognizers employing neutral speech models, e.g. [1–5]. Efforts to improve the performance under LE include design of robust features [3, 4], equalization methods [5] and style-dependent training of acoustic models [1]. Condition-dependent training or design of robust features often results in a decrease of performance when the conditions change [1, 4]. This suggests addressing each of the conditions by a separate dedicated subsystem and implementing a switching mechanism – a condition classifier. Similar idea was proposed successfully in [6], where automatic neural network (ANN) talking style classifier was used to weight outputs of a codebook of style-dependent HMM recognizers. In [7], style classification and speech recognition were performed simultaneously by an N-channel HMM. To each speaking style one HMM dimension was allocated. The approach allowed for style classification on the HMM-state level.

In this paper, a two-stage approach is proposed, using style classifier + independent neutral/LE recognizers. In the first stage, the utterances are classified on the speaking style and in the second stage they are passed to the corresponding dedicated recognizer.

The paper is organized as follows. First, a set of selected features is tested on discriminability in the neutral/LE classification. Several possible setups are compared, the best of which yields the final classification feature vector (CFV). Subsequently, the CFV is used for training ANN and GMM based classifiers. Second, common speech features (MFCC, PLP), special LE-robust features (Expolog [3]) and newly proposed front-end modifications are tested in the neutral/LE digit recognition task, sharing a common back-end architecture. Robustness to changes of average utterance pitch and to emulated noisy backgrounds at various SNRs is compared. Finally, the TSR is designed, employing the style classifier and two recognizers, each using the best performing features found. All the presented experiments were carried out on the CLSD'05 database [8]. The database comprises recordings of Czech neutral speech and Lombard speech uttered in the simulated noisy conditions. In the latter case, a car noise of 95 dB SPL was presented to speakers by closed headphones, yielding high SNR of the recorded speech.

2. Classifying neutral/LE speech

Based on previous studies, only the features providing significant style discriminability on the phoneme/gender-independent level were selected for the neutral/LE classification: vocal intensity, spectral slope of the voiced speech segments and mean and standard deviation of the fundamental frequency. Several frequency bands for spectral slope extraction are considered as well as linear and semitone fundamental frequency representations. Variants with superior discriminability are included in the CFV to train ANN and GMM classifiers. Analyses of feature distributions and training of classifiers were carried out on the *development* set comprising digits and phonetically rich sentences uttered by 8 female and 7 male speakers. *Open* test set comprised digits and sentences uttered by 4 male and 4 female speakers (disjunct from the development ones).

2.1. Features for neutral/LE classification

Voice intensity, spectral slope and fundamental frequency (F_0) are extracted and averaged within the utterance, i.e. each utterance is parameterized by one mean feature vector. For the F_0 feature, also its standard deviation is included in CFV.

Subsequently, distributions of individual features from CFV were obtained by plotting all samples of the particular feature found in *development* data, separately for neutral and LE speech. Neutral/LE discriminability of that feature was then rated based on overlaps of its normalized distributions for neutral and LE speech.

Vocal intensity – since the level of the background noise was almost constant during the CLSD'05 recording, voice intensity changes are displayed directly in the changes of SNR. Normalized distributions of neutral and LE utterance SNRs are shown in Fig. 1. 'Dev' denotes development set, N neutral data, M male and F female utterances.

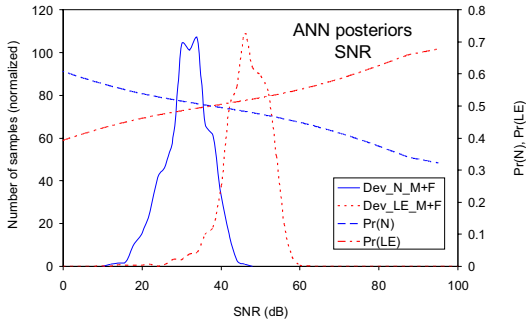

Figure 1: Normalized distributions of SNR, overlap 10.4 %.

Spectral slope – a regression line was fit to the amplitude spectra (initially 0–4k Hz) of the voiced segments to determine the slope, following [10]. In the preliminary study, spectral slopes were analyzed in digit utterances separately for 5 Czech vowels /a/, /e/, /i/, /o/, /u/. Vowel boundaries were determined by means of forced alignment. The mean slopes were observed to differ across vowels, genders and talking styles. Male slopes were in general steeper both on the phoneme and talking style level. In both genders and all vowels, the slopes were steeper for neutral speech, confirming observations in previous works. Female vowel slopes are shown in Tab. 1. '#N, #LE' denotes number of neutral and LE phoneme realizations, respectively.

Set	Females					
	Neutral			LE		
Vowel	# N	Slope (dB/oct)	σ (dB/oct)	# LE	Slope (dB/oct)	σ (dB/oct)
/a/	454	-6.8	1.1	350	-3.2	1.8
/e/	1064	-5.6	1.1	840	-3.1	1.4
/i/	509	-5.0	1.2	405	-2.5	1.8
/o/	120	-8.0	0.9	90	-4.5	1.6
/u/	102	-6.1	0.8	53	-3.9	1.6

Table 1. Mean spectral slopes in female digit vowels.

Subsequently, impact of frequency band-limiting on spectral slope discriminability was evaluated, both gender dependent and independent – see Tab. 2. Results are based on all voiced segments in digits and sentences. 60 Hz high-pass was used to suppress F_0 sub-harmonics. Range 1k–5k Hz covers formants F_2–F_4. Spectral slopes extracted from this band overlap completely and do not discriminate.

Set	Band (Hz)					
	0–8k	60–8k	60–5k	1k–5k	0–1k	60–1k
M	26.0	28.1	29.5	100.0	27.8	28.0
F	26.2	29.0	16.8	100.0	25.8	22.2
M+F	28.1	30.5	29.5	100.0	27.5	26.0

Table 2. Efficiency of various bands for spectral slope based classification – distributions overlap (%).

To exploit the 'discriminative' part of the spectrum, bands were further limited to 0–1k Hz and 60–1k Hz, providing the lowest overlap. Distributions of the slopes obtained from 60–1k Hz band are shown in Fig. 2.

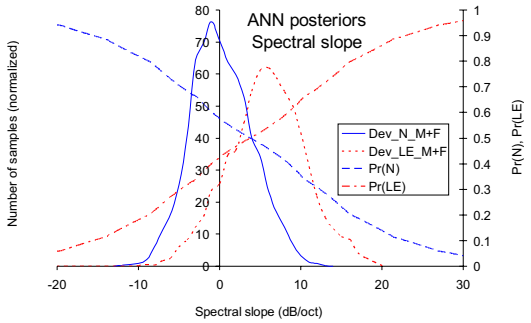

Figure 2: Normalized distributions of spectral slopes extracted from band 60–1kHz; overlap 26.0 %.

Fundamental frequency – distributions of F_0 and F_0 standard deviation are shown in Fig. 3.

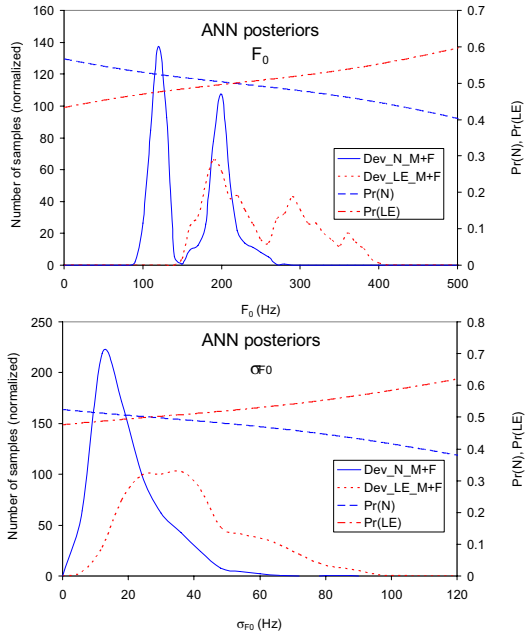

Figure 3: Normalized distributions of F_0 and F_0 standard deviation; overlaps 27.2 and 32.2 %.

2.2. ANN and GMM classifiers

ANN classifier – a fully-connected three-layer MLP (Multi-Layer Perceptron) [13] was trained to estimate posterior probabilities of two complementary classes (neutral/LE speech). The MLP topology was N x 3000 x 2 neurons (input x hidden x output) with sigmoid nonlinearity in the hidden layer and softmax nonlinearity in the output layer. N was set to 1 in the initial single-feature and 4 in the CFV experiment. The MLP was trained against hard targets (the required outputs were either 0 or 1). Since there were 2472 training examples and about 20k trainable parameters, the risk of over-fitting was reduced by using 90% of the data for MLP training and 10% for cross-validation (CV). The final weights were chosen from the epoch achieving maximal accuracy on the CV data. Typically, two iterations were needed. MLP size and learning rate were optimized on the CV data.

GMM classifier – comprises two four-dimensional Gaussian mixture models (GMMs) for neutral and LE speech. Both GMMs employ one mixture per dimension, i.e. each

parameter in the CFV is modeled by a single Gaussian. F_0 distributions in Fig. 2 suggest using more Gaussians, however no improvement in classification accuracy was observed when splitting each mixture into two. Full covariance matrix is used to capture inter-feature variability.

First, the discriminability of individual features from CFV was evaluated by ANN. A set of ANNs was trained on M+F digits and sentences using separate features from CFV. Transfer functions between the input feature value and output posterior probability for neutral – Pr(N) or LE – Pr(LE) classes were evaluated, see Fig. 1, 2. Of some features, two variants were compared: spectral slope was extracted either from 0–8k Hz or from 60–1k Hz; F_0 and σF_0 were considered either on linear scale (Hz) or on log scale (%), see Tab. 3. Confirming observations in Sec. 2.1, $SL_{60\text{-}1kHz}$ provides lower utterance classification error rate (UER) compared to $SL_{0\text{-}8kHz}$. Linear versions of F_0 outperform log ones, presumably because log scale compresses the variations.

Set	SNR_{dB}	$SL_{0\text{-}8kHz}$	$SL_{60\text{-}1kHz}$	F_{0Hz}	σF_{0Hz}	$F_{0\%}$	$\sigma F_{0\%}$
Train	10.9	24.0	19.3	23.8	25.8	25.5	36.3
CV	12.2	20.4	18.2	18.5	25.6	24.8	31.1

Table 3. Efficiency of single feature trained ANN classifier, UER (%). SL – spectral slope. Train set – 2202, CV set – 270 utterances.

Based on this knowledge, for the final CFV were chosen these features: SNR_{dB}, $SL_{60\text{-}1kHz}$, F_{0Hz} and σF_{0Hz}. Performance of ANN and GMM classifiers on the closed and open set is shown in Tab. 4. UER scores are accompanied with 95% confidence interval. ANN classifier outperformed GMM in the open test. It can be attributed to the different nature of classifiers. While GMMs build the models in terms of maximum likelihood, ANNs are trained discriminatively. Note that the training set comprised similar number of samples from both neutral and LE classes.

Set	ANN			GMM	
	Train	CV	Open	Train	Open
# Utter	2202	270	1371	2472	1371
UER (%)	9.9 8.7-11.1	5.6 2.8-8.3	1.6 0.9-2.3	6.6 5.6-7.6	2.5 1.7-3.3

Table 4. CFV-based classification; closed/open test, (M+F).

3. Features for speech recognition

MFCC and PLP features, LE-robust Expolog features [3] and recently proposed features 20Bands-LPC, RFCC-LPC – derived from PLP, and RFCC-DCT – derived from MFCC by modifying filter banks, were compared using HMM-based recognizer. 20Bands-LPC replaces filter bank in PLP by 20 rectangular filters spread over 0–4k Hz, without overlap. RFCC-DCT and RFCC-LPC replace the filter bank in MFCC and PLP, respectively, by the 'repartitioned filter bank' obtained in a data-driven design, see [4, 5] for details. Interestingly, in previous ASR experiments [5], all the above mentioned features outperformed MFCC and PLP in clean (high SNR) LE conditions. This paper extends the study by modified MFCC-LPC and PLP-DCT features (altered cepstral extraction) and by Big1-LPC features (derived from 20Bands-LPC by merging first three bands together). Big1-LPC aims to address the observations that suppressing lower frequency components (up to around 600 Hz) improved LE, while deteriorated neutral speech recognition performance. Merging the first three bands into one reduces the impact of that region on features, while still preserving some information relevant for the neutral speech processing. As significantly higher corruption of recognition performance by LE was observed on female than male speech, the newly proposed features in [4] focused exclusively on females. Hence, for consistency, the following experiments were carried out only on female utterances. The HMMs were trained on 37 Czech SPEECON [14] 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 used for evaluation. Performance of the considered features is shown in Fig. 4.

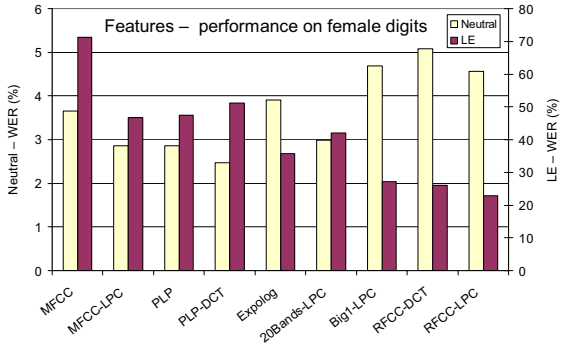

Figure 4: Comparing recognition features – baseline test.

Best results were reached by PLP-DCT and PLP (2.5 %, 2.9 % WER) on neutral speech and by RFCC-LPC on LE speech (23.0 % WER). Subsequently, feature robustness was tested under various levels of added noises from Aurora [15] and Car2E [16] (-5÷25 dB with 5 dB step; infinity dB). Note that the same noises from Car2E were employed, which were used while recording the CLSD database. Moreover, efficiency of full-wave rectified spectral subtraction using Burg's cepstral VAD, as implemented in CTUCopy open source tool [18], was tested, see Tab. 5. 'NSeff' denotes a noise level till which the noise subtraction improved the recognition performance of the best features. Two best performing features are listed for each condition.

Noise	Neutral	LE	NSeff (dB)
Airport	MFCC, 20Bands-LPC	Big1-LPC, RFCC-LPC	None
Babble	MFCC, MFCC-LPC	RFCC-LPC, Expolog	10
Car2E	Expolog, 20Bands-LPC	RFCC-LPC, Big1-LPC	-5
Restaurant	MFCC, 20Bands-LPC	RFCC-LPC, Big1-LPC	-5
Street	20Bands-LPC, MFCC	RFCC-LPC, Big1-LPC	0
Train station	20Bands-LPC, MFCC	RFCC-LPC, Big1-LPC	-5

Table 5. Features performing best on neutral and LE noisy speech.

Subsequently, the dependency of selected features' performance on utterance's average F_0 was evaluated.

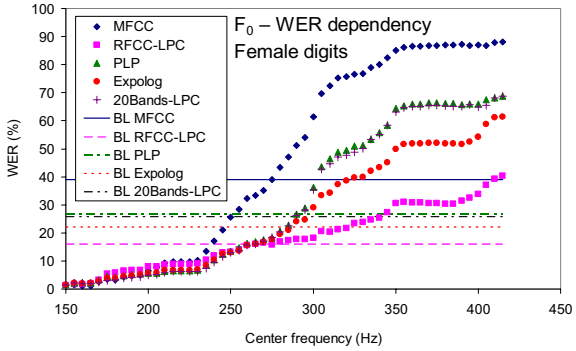

Figure 5: WER(F_c) dependency. BL – baseline WER on the whole merged neutral + LE set.

The neutral and LE digits were merged, yielding a set of utterances with F_0 in range approx. 100–500 Hz. Utterances with F_0 falling within a sliding window of the bandwidth 100 Hz were assigned to the test set F_c (F_c – central frequency of the window). The window was shifted by 5 Hz starting from F_c = 150 Hz. At each step, a recognition was performed to sample WER(F_c) dependency. As shown in Fig. 5., for F_c starting at 250 Hz (F_0 200–300 Hz) RFCC-LPC outperforms the other features, which is consistent with its superior robustness to LE observed in Fig. 4, as LE is displayed in the increase of F_0.

4. Two-stage recognition

PLP and RFCC-LPC features were employed in the neutral and LE dedicated recognizers respectively. PLP represents 'common' features for neutral speech, RFCC-LPC was chosen for its superior properties on female LE speech (see Fig. 4). TSR systems comprising ANN and GMM classifiers followed by dedicated recognizers (of the same architecture as in Sec. 3) were implemented and tested in the open female digits task. Also performance of standalone dedicated recognizers when recognizing both neutral and LE set was evaluated. As shown in Tab. 6., both ANN and GMM TSRs yielded performances of the 'optimal' features on both neutral and LE set, providing and improvement 6.5→4.2 % WER on neutral and 48.1→28.4 % WER on LE compared to the isolated recognizers exposed to the adverse style.

Set		Real neutral	Real LE
# Female digits		1439	1837
WER (%)	PLP	4.3 3.3-5.4	48.1 45.8-50.4
	RFCC-LPC	6.5 5.2-7.7	28.3 26.2-30.4
	ANN Tandem	4.2 3.2-5.3	28.4 26.4-30.5
	GMM Tandem	4.4 3.3-5.4	28.4 26.4-30.5

Table 6. Performance of TSR vs. dedicated recognizers.

5. Conclusions

Efficiency of selected features for neutral/LE classification was evaluated on the CLSD'05 database. Discriminative properties of spectral slope extracted from various frequency bands were studied, finding the band 60–1k Hz superior to others. It was found that linear (Hz) F_0 representation provides better discrimination as compared to log (semitones) one. ANN and GMM gender independent neutral/LE classifiers were trained on the classification feature vector formed by SNR_{dB}, $SL_{60\text{-}1kHz}$, F_{0Hz} and σF_{0Hz}. ANN displayed superior performance on the open task.

Common, special and newly proposed features for speech recognition were compared in neutral/LE digits tasks and tested on robustness to various noises and changes in fundamental frequency of speech. On neutral speech, both features employing DCT or LPC cepstral coefficients were efficient, depending on the type of noise. RFCC-LPC features displayed the best performance on LE female set in all conditions. This confirms the observation that LPC based cepstral coefficients better model Lombard speech spectra even if noise is present [17]. Finally, a two-stage recognition system employing style classifier + neutral (PLP)/LE (RFCC-LPC) recognizers was designed and tested, yielding and improvement 6.5→4.2 % WER on neutral and 48.1→28.4 % WER on LE when compared to the isolated recognizers exposed to the opposite style than for which they were designed.

6. Acknowledgments

The presented work was carried out within the joint project "Normalization of Lombard Effect" of Siemens AG and CTU in Prague. The theoretical part was supported by GAČR 102/05/0278 "New Trends in Research and Application of Voice Technology" and experimental by 1ET201210402 "Voice Technologies in Information Systems."

7. References

[1] Hansen, J.H.L., "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recog.," Speech Comm., 20(2):151-170, 1996.

[2] Junqua, J.C., "The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers," JASA, 93(1): 637-642, 1993.

[3] Bou-Ghazale, S.E., Hansen, J.H.L., "A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech under Stress," IEEE Transactions on SAP, vol. 8, no. 4, pp. 429–442, July 2000.

[4] Bořil, H., Fousek, P., Pollák, P., "Data-Driven Design of Front-End Filter Bank for Lombard Speech Recognition," Interspeech'06, Pittsburgh, pp. 381-384.

[5] Bořil, H., Fousek, P., Sündermann, D., Červa, P., Žďánský, J., "Lombard Speech Recognition: A Comparative Study," 16th Czech-German Workshop on Speech Processing, Prague, pp. 141-148, 2006.

[6] Womack, B., Hansen, J.H.L., "Stressed Speech Classification with Application to Robust Speech Recognition," NATO-ESCA Workshop on Speech Under Stress, pp. 41-44, Lisbon, Portugal, Sept. 1995.

[7] Womack, B.D., Hansen, J.H.L., "N-Channel Hidden Markov Models for Combined Stress Speech Classification and Recognition," IEEE Transactions on SAP, 7(6):668-677, 1999.

[8] Bořil, H., Pollák, P., "Design and Collection of Czech Lombard Speech Database," Proc. Interspeech'05, Lisboa, pp. 1577-1580, 2005.

[9] Cummings, K., Clements, M., "Analysis of Glottal Waveforms across Stress Styles," Proc. ICASSP 1990, 1:369-372, April, 1990.

[10] Summers, V. Johnson, K., Pisoni, D. and Bernacki, R., "Effects of Noise on Speech Production: Acoustic and Perceptual Analyses", JASA 84:917-928, 1988.

[11] Lane, H. et al., "Regulation of voice communication by sensory dynamics," JASA, 47:618-24, 1970.

[12] Titze, I., Sundberg, J., "Vocal intensity in speakers and singers," JASA, 91:2936-2946, 1992.

[13] http://www.icsi.berkeley.edu/speech/faq.

[14] http://www.speechdat.org/speecon.

[15] Hirsch, H.G., Pearce, D., "The AURORA Experimental Framework for the Performance Eval. of Speech Recog. Systems under Noisy Cond.," In ASR-2000, 181-188.

[16] Pollák, P., Vopička, J., Sovka, P., "Czech Language Database of Car Speech and Environmental Noise," EUROSPEECH-99, 5:2263-6, Budapest, Hungary 1999.

[17] Hansen, J.H.L., "Speech Under Stress," Tutorial session, Interspeech'06, Pittsburgh, 2006.

[18] http://noel.feld.cvut.cz/speechlab/start.php?page=download&lang=en.