INTERSPEECH
2009 BRIGHTON

Complementarity of MFCC, PLP and Gabor features in the presence of speech-intrinsic variabilities

Bernd T. Meyer and Birger Kollmeier

Medical Physics, Institute of Physics, University of Oldenburg

bernd.meyer@uni-oldenburg.de

Abstract

In this study, the effect of speech-intrinsic variabilities such as speaking rate, effort and speaking style on automatic speech recognition (ASR) is investigated. We analyze the influence of such variabilities as well as extrinsic factors (i.e., additive noise) on the most common features in ASR (mel-frequency cepstral coefficients and perceptual linear prediction features) and spectro-temporal Gabor features. MFCCs performed best for clean speech, whereas Gabors were found to be the most robust feature in extrinsic variabilities. *Intrinsic* variations were found to have a strong impact on error rates. While performance with MFCCs and PLPs was degraded in much the same way, Gabor features exhibit a different sensivity towards these variabilities and are, e.g., well-suited to recognize speech with varying pitch. The results suggest that spectro-temporal and classic features carry complementary information, which could be exploited in feature-stream experiments.

Index Terms: automatic speech recognition, speech-intrinsic variabilities, feature extraction, spectro-temporal features

1. Introduction

Human listeners outperform automatic speech recognition (ASR) systems not only in acoustically challenging situations (e.g., in the presence of noise or competing talkers), but also when clean speech is to be recognized. Intrinsic factors such as gender, speaking rate and style, dialect, accent, and vocal effort contribute to the vast variability and aggravate finding auditory models that adequately model spoken language. Our work is motivated by the idea to narrow the gap between human and automatic speech recognition by learning from the principles in the human auditory system.

In this study, we used a speech database with speech-intrinsic variabilities to study the effect of speaking rate, effort and speaking style on ASR performance. This corpus was introduced as a tool for man-machine-comparison in speech recognition and contains short non-sense utterances, which avoids high-lexical influence. A hidden Markov model classifier was combined with three different feature types, namely mel-frequency cepstral coefficients (MFCCs), perceptual linear prediction features (PLPs) and Gabor features. MFCCs [1] and PLPs [2] are the most common feature types in ASR and have been compared regarding their robustness towards noise earlier [3]. PLPs were often found to give small improvements over MFCCs, especially in noisy environments or when training and test conditions were not well-matched. However, their robustness against the above-mentioned intrinsic parameters has not been studied so far. Gabor features are physiologically motivated features, and are based on Gabor filters which are a simple model of the spectro-temporal receptive fields in the primary auditory cortex [4]. It was investigated if the explicit use of spectro-temporal information helps to increase overall robustness against extrinsic and intrinsic factors.

2. Speech database

The Oldenburg Logatome corpus (OLLO) [5] was used to analyze the influence of variabilities on ASR performance. It consists of non-sense utterances (i.e., combinations of vowel-consonant-vowel (VCV) and consonant-vowel-consonant (CVC) with the outer phonemes being identical) spoken with different speaking rates and efforts. The database has been used before for studies on the effect of dialect both on human and automatic speech recognition and is freely available for research purposes at http://sirius.physik.uni-oldenburg.de. OLLO contains 14 central consonants (/p/, /t/, /k/, /b/, /d/, /ɡ/, /s/, /f/, /v/, /n/, /m/, /ʃ/, /ʦ/, /l/) and ten central vowels (/a/, /aː/, /ɛ/, /e/, /ɪ/, /i/, /ɔ/, /o/, /ʊ/, /u/).

40 German speakers from four dialect regions and ten speakers with a French accent were recorded for the database. This study focuses on effects induced by different speaking efforts and rates. Hence, only speech from non-dialect speakers (originating from the north-western part of Germany) was used for the ASR experiments (cf. Section 3.3.1).

2.1. Speech-intrinsic variabilities

The choice of variabilities for the corpus was based on ASR experiments with annotated test corpora that compared the performance of automatic recognizers in the presence or absence of these variabilities. The largest impact on performance was observed for varying speaking rate, speaking style, speaking effort, and dialect/accent. Each logatome was recorded in normal (or neutral) speaking style as a reference. In addition, each of the five selected variabilities (i.e., fast and slow speaking rate, loud and soft speaking style, and question which refers to rising pitch) was recorded. To provide a broad test and training basis for ASR experiments and to allow for an analysis of intra-individual differences, each logatome was recorded three times which resulted in 150 × (5+1) × 3 = 2,700 logatomes per speaker.

3. Feature types

3.1. Cepstral coefficients and perceptual linear prediction features

MFCCs [1] and PLPs [2] are the most common methods in feature extraction; both encode the smoothed short-time Fourier transform (STFT) magnitude, which is typically computed every 10ms using an overlapping analysis window of 25ms. For the computation of MFCCs, a pre-emphasis is applied to the signal before calculating the STFT. Each frame is then processed by a mel-filterbank (which approximates the response of the human ear), compressed with the logarithm and transformed to cepstral parameters using an inverse discrete cosine transformation. By selecting several (typically 12 or 13) lower cepstral coefficients, only the coarse spectral structure is retained. This processing results in mostly decorrelated features. PLP features incorporate further psychoacoustic constraints: Linear prediction coefficients are computed from a perceptually weighted nonlinearly compressed power spectrum. The power spectrum is obtained with a Bark filterbank with a subsequent equal-loudness pre-emphasis and a compression based on Steven's power law (i.e., values are compressed by cube-root). The linear prediction coefficients are then transformed to cepstral coefficients. In summing up, MFCCs and PLPs are similar in many aspects, but their differences (e.g., mel- vs. Bark-scaling of the filterbank, compression based on the logarithm vs. Steven's law, differerent pre-emphasis schemes) were found to yield different levels of robustness: MFCCs have been reported to perform very well for recognition of clean utterances or when there is no significant mismatch between training and test noise, while PLPs are often preferred when training and test do not match.

For the presented experiments, both feature types were calculated using the rastamat Matlab toolbox [6] with parameters that resemble feature extraction from the HTK software [7], i.e. the filter bank used 20 frequency channels; the 13-dimensional features were concatenated with delta and acceleration coefficients. Signals with 16kHz bandwidth were used as input to the front-ends.

3.2. Spectro-temporal Gabor features

Gabor features are motivated by physiological measurements in the primary auditory cortex (A1) of several mammal species. Spectro-temporal Gabor filters, which serve as a simple model of spectro-temporal receptive fields of neurons in A1, were proposed for feature extraction in ASR by Kleinschmidt and Gelbart and have been shown to increase the robustness towards extrinsic variabilities [4]. The advantage over the MFCC baseline was most striking for mismatched training and test SNRs and noise signals. However, in this study we focus on the robustness against intrinsic variations and use a paradigm with matched training and test conditions.

The features were calculated by processing a spectro-temporal representation of the input signal by a number of 2-D modulation filters, as depicted in Fig. 1. The filtering was carried out by performing a 2-D correlation of the input representation with each filter function and a subsequent selection of the desired frequency channel of the output. This yielded one output value per frame and filter. Log mel-spectrograms served as input for the filter process.

The two-dimensional complex Gabor function $G(n, k)$ is defined as the product of a Gaussian envelope $g(t, f)$ and the complex sinusoidal function $s(t, f)$. We substituted the Gaussian with a Hanning envelope $h(t, f)$, which resulted in improved results on a digit recognition task [8]. The envelope width is defined by the window lengths W_t and W_f, while the periodicity is defined by the radian frequencies ω_t and ω_f with t and f denoting the time and frequency index, respectively. The two independent parameters ω_t and ω_t allow the Gabor function to be tuned to particular directions of spectro-temporal modulation, including *diagonal* modulations. A further parameters is the center of mass of the envelope in frequency f_0.

The window length was chosen depending on the modulation frequency ω_x, respective the corresponding period T_x, either with a fixed ratio $\nu_x = T_x/2\sigma_x$ = 1 to obtain a 2D wavelet prototype or by allowing a certain range $\nu_x = 1..3$ with individual values for T_x being optimized in the automatic feature selection process. From the complex results of the filter operation, real-valued features were obtained by using the real, imaginary or absolute part only. Special cases are temporal filters ($\omega_k = 0$) and spectral filters ($\omega_n = 0$). In these cases, W_x replaces $\omega_x = 0$ as a free parameter, denoting the extent of the filter, perpendicular to its direction of modulation.

A suitable set of Gabor filters was determined with the Feature Finding Neural Network (FFNN), which is a search algorithm based on a linear neural net. Random filters with physiologically motivated parameteter constraints are employed to calculate features, which are subsequently used to train and test the speech recognition performance of the linear classifier. The relevance of each filter is determined by discarding the corresponding feature component from the feature vector, and calculating the increase of error rate without this feature being used. The least relevant filter is replaced by a randomly drawn new one. This process is repeated until the maximum number of iterations is reached. The filter set used in this study has been optimized using a German database containing noisy digits (Zifkom database). Since the optimization was carried out on data sampled at 8kHz, utterances from the OLLO corpus were resampled to 8kHz bandwidth when used as input for the Gabor front-end. For details on the FFNN algorithm, the reader is referred to [8].

In earlier studies, a large increase in recognition accuracy was obtained by using a Tandem recognizer, i.e., the transformation of features with a non-linear neural network (or multilayer perceptron (MLP)). The resulting posteriors were decorrelated using a principal component analysis (PCA), and fed to a hidden Markov model [9]. In this work, we test two variants of Gabor features, namely the original 80-dimensional filter output without delta features and the filter result processed by a Tandem system, as shown in Fig. 1, which yields 56-dimensional feature vectors. The training and forward run of the neural net are carried out as described in [8].

3.3. Classification system

3.3.1. Training and test set

Utterances of the OLLO database (cf. Section 2) from three male and three female talkers without dialect ($\sim$17k speech items) served as training data, logatomes from the four remaining speakers (speaker indices $\{1,2,6,8\}$, $\sim$11k utterances) were used for the test, with speech-intrinsic variations being equally distributed in both sets. The chosen segmentation of the corpus results in a speaker- and gender-independent ASR system.

To study the effect of noise on the different feature extraction schemes and on changes in speaking rate and effort, a stationary, speech-shaped noise was used. Noise was added at SNRs ranging from -10 to 10dB in 5dB-steps, and the noisy speech data was subsequently used to train and test the backend, with identical train and test SNR. The SNR was calculated by relating the root-mean-square (rms) value of the speech segments of each audio signal and the rms value of the masking noise of equal length. A simple voice detection algorithm based on an energy criterion was used to extract connected speech segments. Additionally, the classifier was trained and tested with clean speech.

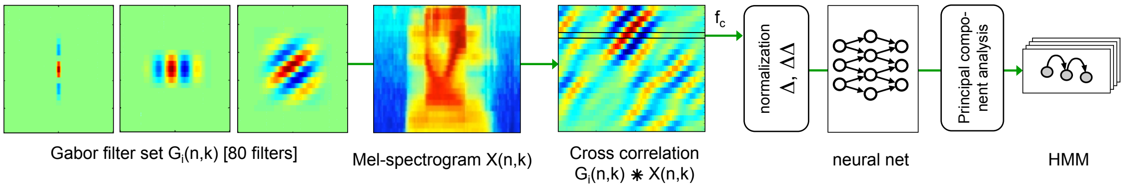

Figure 1: Gabor features are calculated by correlation of each filter with a mel-spectrogram and subsequent selection of the center frequency f_c associated with each filter. This results in 80-dimensional vectors that are used to train and test a Tandem recognizer, which consists of a neural net and an HMM. Gabor functions on the left are examples of purely temporal, spectral and spectro-temporal filters. The cross-correlation in this example was obtained with a spectro-temporal filter that emphasizes the diagonal transient.

3.3.2. Classification system

ASR experiments were carried out with a Hidden Markov Model (HMM) with three states and eight Gaussian mixtures per state. Logatomes with the same outer phoneme were used to train and test HMMs which were subsequently used to classify the central phoneme in CVCs and VCVs, i.e., confusion occured only between central phonemes. The HTK classifier was used for the experiments [7].

4. Results and discussion

4.1. Overall results

Phoneme recognition rates for the different feature types are presented in Fig. 2. MFCCs result in the lowest error rates for the relatively easy task of recognizing clean speech with a clean-trained recognizer. At SNRs below 0dB, PLPs perform better than MFCCs. The differences between PLPs and MFCCs show a slightly higher robustness of PLPs for this task, supporting the results in [3]. Gabor features delivered better average performance when being used with a Tandem recognizer (which includes a PCA as final processing stage), compared to directly using them as input for the HMM. A reason for this might be the correlation of Gabor feature components, which is a disadvantage when using diagonal covariance matrices in HMMs, as it has been done here. Spectro-temporal features produce higher error rates than MFCCs and PLPs in clean speech, but at 5 and 10dB SNR, Gabor-MLP features are on par with MFCCs, and perform better below 5dB SNR (with relative reductions in word error rate up to 17% compared to MFCCs).

4.2. Speech-intrinsic variabilities

ASR phoneme error rates depending on speech intrinsic variations are shown in Fig. 3 for three feature types. Original Gabor features produced scores between PLP and Gabor-MLP and are not shown in the figure. Intrinsic variations induce large differences in performance. On average, the reference condition and slow speaking rate result in rather low error rates, while the conditions 'fast' and 'soft' yield an increase of errors. As an example, MFCCs errors increase by over 10% for fast spoken utterances compared to the category 'normal'. Similar differences are observed for the other features. However, MFCCs and PLPs seem to be similarly affected by the intrinsic variations: Throughout all variabilities, MFCCs perform better than PLPs for the 5dB-SNR task, and the average results are quite similar. We conclude that the differences between feature extraction schemes for MFCCs and PLPs do not influence their sensivity towards these intrinsic variabilities. Gabor-MLP features show the best average performance throughout all variabil-

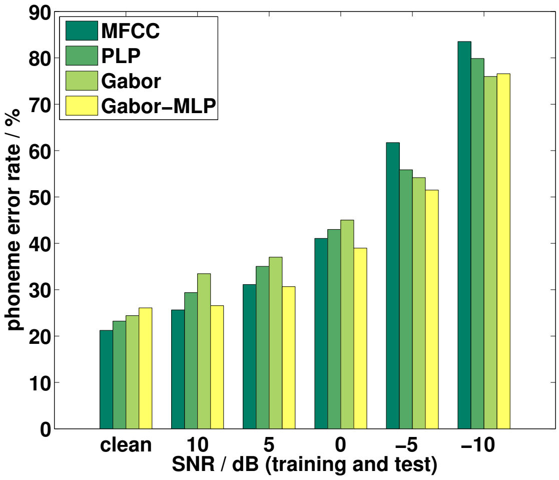

Figure 2: Overall error rates for different feature extraction schemes. Results were obtained by training and testing the ASR system with the same signal-to-noise ratio.

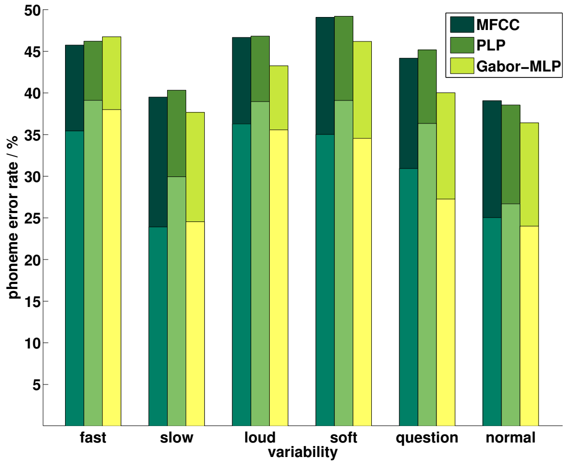

Figure 3: ASR phoneme errors for different feature types, depending on speaking style, rate and effort. Scores were obtained by training and testing the recognizer with logatomes at 5dB SNR (light bars) which resulted in similar overall performance for MFCC and Gabor-MLP features. Errors averaged over all SNRs from -10dB to 10dB and clean speech are depicted by darker bars.

ities, with the exception of fast speaking style. This condition also shows relatively weak performance for the 5dB-SNR task. On the other hand, low error rates are obtained for the condition 'question'. Thus, the usage of spectro-temporal features is not only beneficial for overall performance, but also results in different sensivity towards several intrinsic variations. The reason for the high error rates for fast spoken utterances might be that the optimization of the filter set was carried out on German digits that were spoken at normal speaking rate. Higher spectro-temporal modulation frequencies, which might be better suited to detect, e.g., formant transitions of speech at high speaking rate, may therefore not be included in the filter set.

4.3. Information transmission

The acoustic cues important for consonant identification are analyzed by decomposing phonemes into their linguistic or articulatory features (AFs). This method of data analysis is based on works by Miller and Nicely [10] who proposed several AFs to group speech stimuli. The amount of transmitted information (TI) associated with each feature is calculated based on the confusions of these features.

Fig. 4 shows the normalized transmitted information for consonant and vowel phonemes, as well as for several articulatory features. MFCCs exhibit very good performance for the overall classification of consonants and for the consonant features place and manner, while Gabor-MLPs show the highest TI scores for the voicing feature. The reduced performance of MFCCs and PLPs for voicing presumably arises from the spectral smoothing and elimination of fine structure. Gabor-MLPs perform best for the recognition of vowels and vowel-associated articulatory features. In previous studies, Gabor features have been shown to carry complementary information compared to MFCCs [8]. The observed differences in Figs. 3 and 4 indicate that this complementarity also applies to intrinsic variations and articulatory features, which motivates further experiments that combine properties of spectro-temporal features with MFCCs and PLPs.

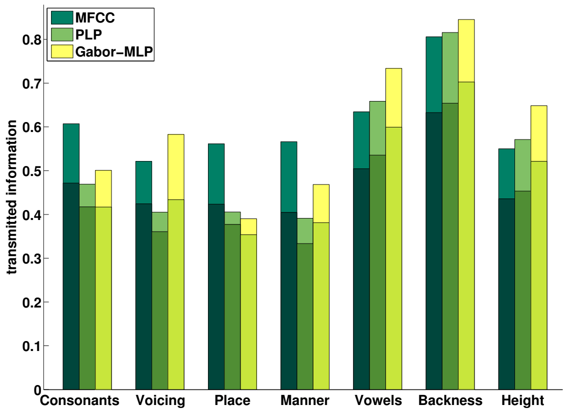

Figure 4: Transmitted information based on consonant and vowel confusion matrices, and for the articulatory features manner, place and voicing (derived from consonant confusions) and height and backness (based on vowel confusions). As in Fig. 3, the lighter bars denote results obtained with training and test at 5dB SNR, while darker bars show results averaged over all SNR conditions.

5. Conclusions

In this work, we analyzed the properties of different ASR features towards extrinsic and intrinsic variabilities. MFCCs were found to produce best results in acoustically optimal conditions with matched train-test conditions, while PLPs are better-suited for phoneme recognition below 0dB SNR than MFCCs. Error rates obtained with Gabor features at low SNRs were 17% lower than with MFCCs, which documents the robustness towards extrinsic factors.

Intrinsic variations (speaking rate, style and effort) had a strong impact on ASR performance. However, MFCCs and PLPs seem to be equally affected by these variations. On the other hand, our analysis showed that Gabor features differ from MFCCs and PLPs regarding sensivity towards intrinsic parameters, since utterances spoken as question produced relatively low error rates, while fast spoken utterances were better recognized with MFCCs. The analysis based on transmitted information showed that voicing and vowel-associated features are better encoded by Gabor features, whereas the highest TI scores for place and manner of articulation were found for MFCCs. We therefore argue that Gabor features and MFCCs/PLPs carry complementary information, which could be exploited in feature-stream experiments.

6. Acknowledgments

Supported by the DFG (SFB/TR 31 'The active auditory system'; URL: http://www.uni-oldenburg.de/sfbtr31). We thank Thomas Brand, Jörg-Hendrik Bach, and Birgit Lügering for their support and contributions to this work.

7. References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[3] P. Woodland, M. Gales, and D. Pye, "Improving environmental robustness in large vocabulary speech recognition," in *Proceedings of ICASSP*, 1996, pp. 65–68.

[4] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction," in *Proc. ICSLP*, 2002.

[5] T. Wesker, B. Meyer, K. Wagener, J. Anemüller, A. Mertins, and B. Kollmeier, "Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines," in *Proceedings of Interspeech*, Lisbon, Portugal, 2005, pp. 1273–1276.

[6] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource. [Online]. Available: http://www.ee.columbia.edu/ dpwe/resources/matlab/rastamat/

[7] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.2)," Cambridge University, Eng. Dept., 2002, techn. Report.

[8] B. Meyer and B. Kollmeier, "Optimization and evaluation of Gabor feature sets for ASR," in *Proc. Interspeech*, 2008.

[9] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000.

[10] G. Miller and P. Nicely, "An analysis of perceptual confusions among some english consonants," *J. Acoust. Soc. Am.*, pp. (2) 338–352, 1955.