Comparison of speech parameterization techniques for Slovenian language

Rok Gajsek and France Mihelič

Abstract— **The goal of speech parameterization is to extract the relevant information about what is being spoken from the audio signal. In modern speech recognition systems mel-frequency cepstral coefficients (MFCC) or perceptual linear prediction coefficients (PLP) are the two main techniques used. MFCC method is known to give better results when audio recordings are of high quality (no background noise, quality microphone) whereas the PLP performs better when the quality of audio is poor. In an attempt to close the gap between the two methods some modifications to the original PLP method are presented. They are mainly based on using a modified mel-filter bank with a number of filters resembling the number of spectral coefficients.**

In our work the effectiveness of proposed changes to PLP (RPLP features) were tested and compared against the MFCC and original PLP acoustic features. A number of 3-state HMM acoustic models were build using different acoustic feature setups (different filter banks, different number of filters) in order to assess which parameterization technique gives superior recognition accuracy. To achieve a more robust estimate of the recognition results when using various parameterizations three databases of different audio quality were used.

I. INTRODUCTION

Speech parameterization is an important step in speech recognition systems. Its aim is to extract from the audio signal the information about the voices (phonemes) that are being spoken. Figure 1 presents an overview of a speech recognition system. Since speech parameterization is the first step in processing the audio input it effects heavily on all other procedures that follow. Hence, it has a significant impact on final recognition accuracy.

Result of speech parameterization is a number based presentation of a short (10 to 50 ms) time period of audio signal. Mainly, two acoustic features are found in today's state of the art speech recognition systems: mel-frequency cepstral coefficients (MFCC) [1] or perceptual linear prediction coefficients (PLP) [2]. MFCC method is known to give superior results when audio recordings are of higher quality (controlled environment, no background noises, quality microphone) whereas the PLP method prevails if recordings are of poorer quality [4]. PLP features are also more robust resulting in better recognition rates when there is a big mismatch between training and testing data [3].

Regardless of the fact that the two methods were developed independently there are many similarities between the two. The differences however lie in the shape of the filter bank, intensity to loudness conversion, equal loudness pre-emphasis and the usage of linear prediction in PLP. It would be very useful to combine the positive aspects of both methods and derive a new speech features that would give superior results regardless of the quality of audio recordings. In the following article we present the proposed modifications of PLP method [4] and apply them to the Slovenian speech. The results will be compared with the MFCC and standard PLP acoustic features to assess if the increase of performance is noticed recognition of Slovenian language also.

Voicetran [5] speech database that consists of three different smaller databases will be used for the tests so that a robust estimate of the effectiveness of the new features can be acquired.

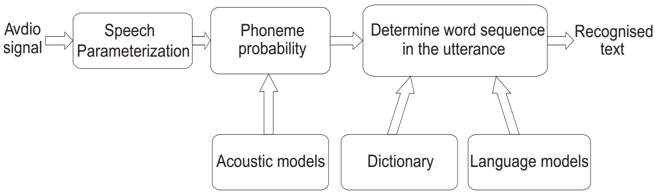

Fig. 1. Speech recognition system overview

II. MFCC ACOUSTIC FEATURES

MFCC acoustic features have been developed some time ago, but they still provide one of the best ways to parameterize speech signals and are thus found in many modern recognition systems. The method is based on Fourier transformation and discrete cosine transformation (DCT). The left part of the figure 2 presents the steps used to calculate the MFCC features.

A. Pre-emphasis

The aim of pre-emphasis is to even the spectral energy envelope by increasing the high frequency components in the signal. In the case of MFCCs the pre-emphasis is applied to the speech signal before the short term spectral analysis. It is implemented by using a first order discrete system:

$$H(z) = 1 - a(z{-}1) \qquad (1)$$

where the value of a is between 0.9 and 1.0 (in our work we used $a = 0.97$).

B. Hamming window

In general the speech signal is not a stationary stochastic signal, but if we split the signal in small enough pieces it can be regarded as such. Windowing is precisely that, it splits the signal in smaller (usually 10-50 ms) parts. It is also important that there is an overlap between the windows since we would have a loss of information at the borders. These overlap is

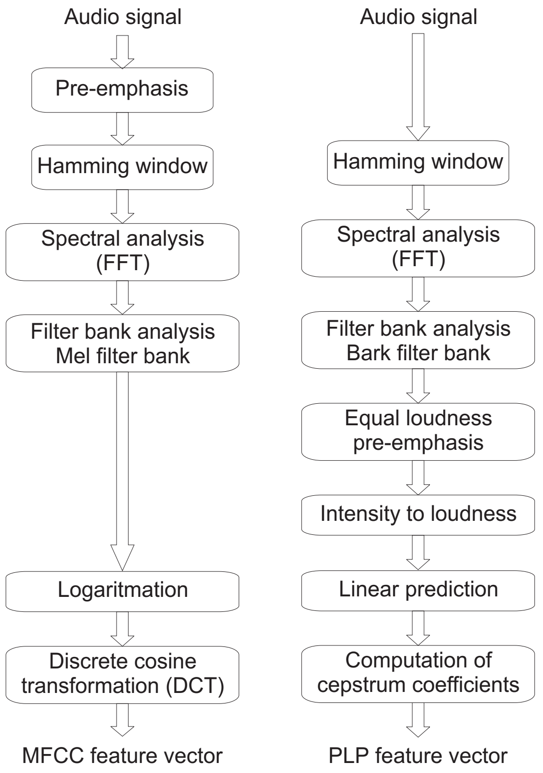

Fig. 2. Overview of MFCC and PLP methods

usually around 50%. Many different windowing functions exist (shown in the figure 3) but in speech processing a Hamming window is mostly used. The Hamming window is defined with the following equation:

$$w(n) = 0.54 - 0.46cos(\tfrac{2\pi n}{N-1}), \quad (2)$$
$$0 \leq n \leq N-1.$$

C. Spectral analysis

Analysis of the human hearing organs have shown that the sound wave is split according to the frequencies. The same thing is accomplished in acoustic feature extraction with short term Fourier transform (discrete Fourier transform is used 3).

$$X(k) = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn} \quad (3)$$

The result of this transformation is the power spectrum of the windowed signal.

D. Filter bank analysis using Mel filter bank

As mentioned in previous chapter the human ear divides the sounds according to the frequencies, but the resolution throughout the frequency scale is not linear (the resolution is smaller at low frequencies and gets bigger as the frequency increases). These behaviour is simulated with the use of filter banks. Many different types of filter banks exists but for MFCC features the Mel filter bank [1] is used. The filters have a triangular shape and following equation (4) defines the centre frequencies of these filters.

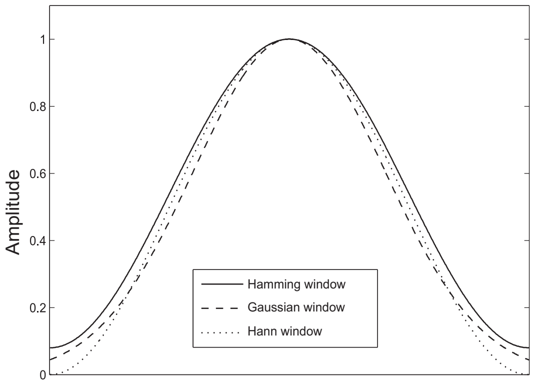

Fig. 3. Comparison of different windowing functions

$$M(k) = 1127 * \log{(1 + \tfrac{f_k}{700})} \quad (4)$$
$$f_{min} \leq f_k \leq f_{max}$$

The start of k filter is at the centre frequency of the $k-1$ filter and the end is at the centre frequency of $k+1$ filter. These way the width of the filter depends on the total number of filters selected (usually around 20 are used), but the overlap between the neighbouring filters is always 50% which can be observed in the figure 4.

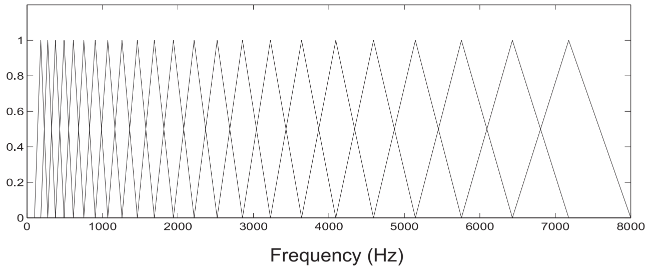

Fig. 4. Mel filter bank

All the coefficients that fall into filter k are first weighted with the value of the filter at the corresponding frequency and then summed. These way the result of filter bank analysis is one value for each filter.

E. Logaritmation

By applying the logarithm function we are again simulating human ear where the sense of volume is not linear. Plus there is another positive aspect of logarithm: the product becomes a sum. If there is a source of noise in the audio signal and if we can define that noise it can be easy deducted after logaritmation.

F. Applying discrete cosine transformation

With the calculations described so far we get a highly correlated set of features. This means that by using statistical models (in speech recognition Hidden Markov Models (HMM) are used mainly) we would have to use full covariance matrix which would significantly increase the computational load. With the use of discrete cosine transformation

(DCT) the features are decorrelated and we can later use diagonal covariance matrices.

$$MFCC_i = \sum_{k=1}^{N} X_k \cos[i(k-\tfrac{1}{2})\tfrac{\pi}{N}] \qquad (5)$$
$$i = 1, 2, ..., M$$

With the step of DCT (equation (5)) a final vector of MFCC coefficients is acquired.

III. PLP acoustic features

The perceptual linear prediction features (PLP) presented at the right side of the figure 1 were developed by Hynek Hermansky [2]. The PLP method is mainly based on the findings from the research of psychoacoustics.

As can be seen from the figure 1 some steps of computing PLP coefficients are the same as with the MFCCs, therefore in the following sections an overview of those that differ is given.

A. Filter bank analysis using Bark filter bank

The application of filter bank is similar for PLP as it is for MFCC only instead of Mel filter bank a Bark filter bank [2] is used. The centre frequencies for the corresponding filters are calculated according to the bark scale (6).

$$\Omega(k) = 6\log\left[\frac{f_k}{600} + \sqrt{\frac{f_k}{600}^2 + 1}\right] = 6\sinh^{-1}\left(\frac{f_k}{600}\right) \quad (6)$$
$$f_{min} \leq f_k \leq f_{max}$$

The shape of the filters is defined by equation (7).

$$C_k(\omega) = \begin{cases} 10^{1.0(\Omega-\Omega_k+0.5)} & , \Omega \leq \Omega_k - 0.5 \\ 1 & , \Omega > \Omega_k - 0.5 \\ & , \Omega < \Omega_k + 0.5 \\ 10^{-2.5(\Omega-\Omega_k-0.5)} & , \Omega \geq \Omega_k + 0.5 \end{cases} \quad (7)$$

$C_k(\omega)$ is a weight of the k filter at frequency ω

Ω_k is a centre frequency of the filter k

$k = 1, 2, \ldots, K.$

The number of filters used is similar to MFCC (around 20), the major difference lies in the shape of the filters which can be observed from figure 5.

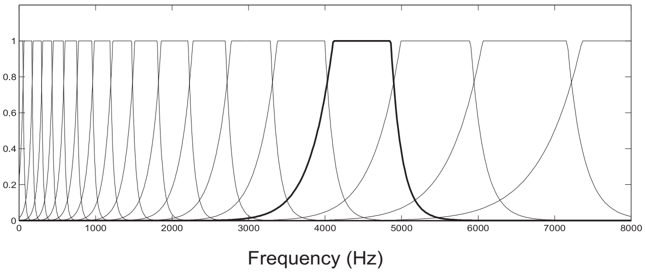

Fig. 5. Bark filter bank

B. Equal loudness pre-emphasis

As oppose to MFCC method where the pre-emphasis is done directly on the speech signal prior to spectral analysis, with PLP the pre-emphises is applied to the short term power spectrum. The proposed function is

$$E(f) = \frac{((f^2 + 1.44 \cdot 10^6)f^4}{(f^2 + 1.6 \cdot 10^5)^2(f^2 + 9.61 \cdot 10^6)}$$

The author also suggests that the first and the last filter should be duplicated since the filters with 0 and $f_{nyquist}$ centre frequencies can not be calculated.

C. Intensity to loudness conversion

Stevens power law describes the relation between the intensity of the audio signal and human perception of loudness. It stats that the perceived loudness is a cubic root of intensity. This is taken into account in calculation of PLP features and a cubic root of every filter bank values is calculated.

D. Linear prediction

Linear prediction means that we try to find the coefficient of the hypothetical signal, whose power spectrum matches the one that is calculated from the windowed audio signal. Fifth order models are generally used for modelling this hypothetical signal. Spectrum of these model has one or at most two spectral peaks that represent the first, or the first and the second formant frequency. The first peak represents the first formant frequency and the second formant frequency is represented with the second peak only if the peaks are at least 4 barks apart. Otherwise both peaks are united into one and positioned in the middle of their original locations. This way we lower the dimension of the spectrum and in some cases increase the frequency resolution since we are not limited anymore with the central frequencies of filters.

E. Computation of cepstrum coefficients

By taking a logarithm of the spectrum of the predicted model and then inverse Fourier transform the parameters of the model are transformed to cepstrum coefficients.

IV. Proposed modifications of PLP

In the article [4] the authors present some modifications of the standard PLP feature extraction method. These changes are based on a study of differences between MFCC and PLP parameterizations. The proposed modifications are presented in the following section.

A. Filter bank analysis

As described above, different filter banks are used in spectral analysis of MFCC or PLP computations. The only major difference between the Mel and Bark filter bank is the shape of the filters, thus the performance of both filter banks is similar [6]. With substitution of Bark filter bank with Mel filter bank in calculation of PLP no increase of performance is observed.

The free parameter in Mel filter bank as noted above is the number of filters. By increasing the number of filters they become narrow but with a small number of filters the loss of information is introduced. A new filter bank is presented where the width of the filters is fixed to 226 Mel and the number of them is equal to the number of spectral coefficients (in our case we used 257 coefficients). Hence, the overlap between the filters is much greater than 50%.

B. Equal loudness pre-emphasis

In MFCC the pre-emphasis is applied to the speech signal before the short term spectral analysis whereas in PLP the equal loudness pre-emphasis is applied to the power spectrum. But nevertheless, there are many similarities between the two [4], thus an application of pre-emphasis as used in MFCC in applied to PLP as well.

C. Linear prediction

The effect of duplication of the first and the last filter is also evaluated, since an overemphasise can occur. The results in [4] show that in combination with signal pre-emphasis the recognition rates are higher if filter duplication is omitted.

Following the above modifications a new acoustic features (named RPLP) are derived.

V. Evaluation

Performance of speech recognition of Slovenian language was tested using the new RPLP features and compared to the standard MFCC and PLP. In all tests the energy was also added to the feature vector. The dynamic properties (Δ and $\Delta\Delta$ coefficients) were computed so the final parameterization vector for MFCC consisted of 39 coefficients (12 MFCCs + energy + Δ + $\Delta\Delta$).

A. Speech corpus

The tests were carried out on a Voicetran speech corpus [5] which consists of three databases: Gopolis, VNTV and K211D. Gopolis and VNTV databases were recorded in controlled environment using quality microphone and K211D is a corpus of weather news reports that were aired on Slovenian national television. These way we have a diverse set of recording in order to obtain a more objective estimate of the speech recognition accuracy achieved using different parameterizations.

B. Results

For each parameterization a monophone acoustic model was build. A left-to-right three state HMM was used to represent one Slovene phoneme. The new RPLP feature extraction algorithm has been implemented in HTK environment [7]. Comparison of phoneme recognition rates is shown in the table I.

The new filter bank with fixed filter width and a large number of filters has also been applied and tested with the standard MFCC method. It can be seen (table I) that the recognition accuracy improves slightly for a relative value of 1.1%.

From the table I it can be observed that improvement (2.5% relative increase of recognition accuracy) is achieved with the new RPLP method of parameterization over the baseline MFCC method. Increase in accuracy is not significant which shows that using more filters and thus dividing the spectral range in more smaller fragments does benefit to better represent of phonemes that were spoken.

Acoustic feature	Phoneme recognition rates
MFCC, 24 filters	42.81%
MFCC, 257 filters	43.29%
PLP, 24 filters	42.71%
RPLP, 24 filters	43.46%
RPLP, 257 filters	**43.89%**

TABLE I
Comparison of recognition accuracy using different parameterization techniques

VI. Conclusion

In the article a modified PLP method of speech parameterization (RPLP) has been presented and its effect on recognition accuracy of Slovenian speech has been evaluated. Since this method is derived as a combination of PLP and MFCC acoustic features these are presented first. Then the proposed modifications to PLP are described and finally all parameterizations are evaluated on a corpus of Slovenian speech. We have shown that the new RPLP features increase the accuracy of the recognition by 2.5% relatively according to the standard MFCC which is in accordance with what the results are in other languages. We are able to draw a conclusion that the proposed modifications of standard PLP method increase the accuracy of the speech recognition system.

References

[1] S. Davis, P. Mermelstein, Comparison of parametric representation for monosyllabic word recognition in continuous spoken sentences, IEEE Trans. ASSP, vol 28., no. 4, 1980, pp. 357-366.

[2] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, Journal Acoust. Soc. Amer., vol 87, no. 4, 1990, pp. 1738-1752.

[3] P. Woodland, M. Gales, D. Pye, Improving environmental robustness in large vocabulary speech recognition, Proceedings of ICASSP, vol. 1, 1996, pp. 65-68.

[4] F. Höning, G. Stemmer, C. Hacker, F. Brugnara, Revising perceptual linear prediction (PLP), Interspeech-2005, 2005, pp. 2997-3000.

[5] F. Mihelic, et al., Spoken language resources at LUKS of the University of Ljubljana, International Journal of Speech Technology 6 (3), 2003, pp. 221-232.

[6] X. Huang, A. Acero, H.-W. Hon, Spoken language processing - A guide to theory, Algorithm, and System development, Upper Saddle River: Prentice Hall, 2001.

[7] S. Young, G. evermann, et al., The HTK Book (for HTK version 3.4), Cambridge University Engineering Department, 2006.