Comparison of MFCC and PLP Parameterizations in the Speaker Independent Continuous Speech Recognition Task

Josef Psutka, Luděk Müller and Josef V. Psutka

University of West Bohemia, Department of Cybernetics, Univerzitní 8, 30614 Pilsen, Czech Republic
psutka@kky.zcu.cz, muller@kky.zcu.cz, psutka_j@kky.zcu.cz

Abstract

The authors of this paper wish to contribute to the discussion about an optimal parameterization of speech signals in speech recognition systems. Our experiments deal with a telephone-based speaker independent continuous speech recognition task in which the MFCC and PLP parameterizations were tested and compared. The benefit of an adjustment of the parameters used in the MFCC and PLP parameterizations to the critical bandwidth of hearing [1] was explored and the impact of the number of filters and enumerated parameters to the recognition accuracy was tested. The results of these experiments showed that the MFCC parameterization is less sensitive to satisfying the theory of the critical bandwidth of hearing than the PLP parameterization. Experiments also proved that 5 PLP-cepstral (including derived 5 delta + 5 delta-delta) coefficients do not afford the best results as could be deduced from recent work [2], [3]. However, after optimal setting both parameterization techniques provided almost comparable results.

1. Introduction

One of the most important components of the speech recognition system is the front-end. In recent papers we can find various recommended parameterization techniques or various modification of standard techniques. Only very few works try to appreciate and compare these techniques mutually and experiments with speaker independent continuous speech recognition are quite rare. On the other hand we can observe the “over-dimensioned” or more sporadically “under-dimensioned” front-ends which respect neither the application tasks nor the real working conditions as time and memory demand.

This paper deals with the MFCC and PLP parameterization techniques. It is well known that both techniques try to accommodate the parameter estimation process to the way humans hear and how they perceive sounds with various frequencies. The concept of critical band rate and critical bandwidth is frequently applied in speech recognition from this point of view. While the problem area of critical-band rate especially for the PLP parameterization technique was discussed in [4], the following part of this article will deal with the critical bandwidth and its benefit both for the MFCC and PLP parameterization. In numerous tests, these techniques were compared for different numbers of filters distributed in the given frequency band and for different numbers of enumerated parameters. We also investigated the influence of the introduction of delta and delta-delta parameters and the impact of the mean and amplitude normalization of particular coefficients on the accuracy of recognition experiments.

All experiments were performed on a continuous speech database pronounced by 100 speakers over a telephone channel. Because speakers called from various places in the Czech Republic the transfer conditions (e.g. noise, distortion, etc.) were generally slightly different for each call. Only the zero-gram language model was used during recognition experiments to gain a better view of the behavior of a word error rate (WER) caused by an adjustment of the front-end.

2. Speech recognition conditions

The recognition experiments were performed with the recognition engine which is a part of a telephone dialogue system [5] built at the Department of Cybernetics, University of West Bohemia, Pilsen. The recognition engine is based on a statistical approach. It incorporates front-end, acoustic model, language model and decoding block. The basic speech unit of our system is a triphone. Each individual triphone is represented by a three states HMM with a continuous output probability density function assigned to each state. At present we use 8 mixtures of multivariate Gaussians for each state. As the number of Czech triphones is too large, phonetic decision trees were used to tie states of Czech triphones. The digitalization of an analogue telephone signal was provided by a telephone interface board DIALOGIC D/21D at 8 kHz sample rate and converted to the mu-law 8-bit resolution format. The aim of the front-end processor is to convert continuous speech into a sequence of feature vectors. This parameterization process can run either on the Mel-Frequency Cepstral Coefficients (MFCCs) or the PLP coefficients.

The decoder uses a crossword context dependent HMM state network, which is generated by a Net generator. The Input of the Net generator is a text grammar format represented by an extended BNF that respects the VoiceXML description. The whole net consists of one or more connected grammars. The decoder uses a Viterbi search technique with an efficient beam pruning.

Because a variety of noise sounds, e.g. load breath, noise of a telephone channel can appear in an utterance, a set of noise HMM models was introduced and trained in order to capture these noise sounds. The speech material for all experiments was taken from the Czech telephone corpus collected at the Department of Cybernetics. The corpus consists of a read speech transmitted over a telephone channel. One hundred speakers were asked to read 40 sentences. These sentences were selected from Czech newspapers in order to contain the most occurring triphones of the Czech spoken language. The corpus obtained was manually annotated and phonetically transcribed. Then it was randomly divided so that 100 sentences created the test part and the remaining part of the corpus formed the training part. The
vocabulary of our task contained 475 different words. Since several words had multiple different phonetic transcriptions the final vocabulary consisted of 528 items. In all recognition experiments a language model based on a zero-gram was applied. It means that each word from a vocabulary is equally probable as a successor of a given word in the recognized utterance. For that reason the perplexity of the task was 528.

3. Auditory-based front-ends

The MFCC and PLP-based front-ends attempt to model the auditory processing up to activation of the inner hair cells by the basilar membrane vibrations. This simulation is usually performed through a selective model which is implemented by a filter bank whose center frequencies are spaced along the frequency axis satisfying the critical-band scale and whose particular filter widths correspond to the theory of critical bandwidths [2]. The most common critical-band scales are the mel-scale and the bark-scale in which the filters are distributed along the frequency axis approximately linear up to about 1000 Hz and logarithmic above 1000 Hz. The number of critical bands depends on the whole frequency band width and/or the sampling frequency $F_s$. For the telephone frequency band with $F_s=8$ kHz approximately from 15 to 17 critical bands are recommended.

3.1. MFCC parameterization

The computational algorithm of the MFCC parameterization is realized by the bank of symmetric overlapping triangular filters spaced linearly in a mel-frequency axis, according to auditory perceptual considerations. The spacing as well as bandwidth of the particular filters is determined by a constant mel-frequency interval. In our case the spacing was approximately 145 mels and the width of the triangle was 290 mels. So, for telephone frequency band (0–2146 mels) with the sampling frequency $F_s=8$ kHz we obtained (using critical-band concept) about 15 filters distributed up to the Nyquist frequency. The MFCC parameterization was accomplished by the computation of mel-cepstral coefficients $c(1), \ldots, c(M)$. In practice these coefficients are obtained by applying an inverse DFT to the log-energy output of the filter bank. The set of cepstral coefficients is complemented by the coefficient $c(0)$ which approximates the average log-energy of the signal (this coefficient is often replaced by the energy directly computed which approximates the average log-energy of the signal (this

3.2. PLP parameterization

The front-end used in this case was based on the PLP parameterization described in [3]. For transformation of a power speech spectrum to a corresponding auditory spectrum the PLP combines three components from the psychophysics of hearing: the critical-band spectral selectivity, the equal-loudness curve and the intensity-loudness power law. To execute this process we have to perform following steps:

- Computation of short-term speech spectrum
- Nonlinear frequency transformation and critical-band spectral resolution. Modeling of these phenomena is performed in the PLP either by the nonlinear transformation of bandwidths of the consequent in the Bark scale and by the construction of masking curves that simulate critical-band of hearing and are modeled by the band-pass filters. The centers of filters are spaced in the Bark domain linearly with the step approximately 1 Bark. As the speech signal covers the range from 0 to 4 kHz the corresponding range in the Bark scale was 0 to 15.57 Bark. The last ($M-1$)th filter had a center in the value of 0 Bark, the last ($M-1$)th filter was centered in the value of 15.57 Bark.

- Linear-mel frequency transformation and PLP-cepstral representation. The PLP cepstral coefficients $c(1), \ldots, c(Q)$ are computed by the standard approach from the $Q$ PLP predictive coefficients. For the final acoustic modeling we extended the original PLP-cepstral representation with derived delta and delta-delta features. In fact the dimension of the pattern space in which the acoustic models of triphones were built was 3Q.

4. Influence of the front-end adjustment on the recognition accuracy

The goal of these experiments was to explore how the adjustment of the front-end to the critical bandwidth of human hearing influences the recognition accuracy. The second goal was to verify, how many coefficients enumerated in the MFCC and PLP parameterizations afford the best recognition accuracy. These experiments were, in fact, performed either by an increasing the number of bands (filters) for a fixed number of coefficients or on the contrary by increasing the number of coefficients for a fixed number of bands (filters). At the same time the filters were spaced linearly in mel- or bark-scale over the whole frequency band. It means, that the change of the filter number influenced to the bandwidth of individual filters. Further experiments explored on the one hand the favorable effect of acceleration coefficients and on the other hand the positive influence of an amplitude and a cepstral mean normalization of individual coefficients. For an evaluation of recognition results we used the standard measure – the accuracy (Acc) defined as

$$\text{Acc} = \frac{(N - D - S - I)}{N} \times 100\%,$$

where $N$ is the total number of words in the reference transcription, $S$ is the number of substitution errors, $D$ is the number of deletion and $I$ the number of insertion errors.

4.1. Influence of coefficient and filter numbers on recognition accuracy for the MFCC parameterization

A number of experiments was performed in which the bank of filters was increased from 4 to 26 filters and the number of coefficients was enumerated from 4x3=12 to 24x3=72. Let us remind that these sets of coefficients consisted of the static MFCC complemented by the derived delta and delta-delta coefficients. In all experiments the bank of filters was distributed over the whole frequency band 0–2146 mels. It means that in case of 4 filters the spacing of these symmetric overlapping triangular filters was approximately 430 mels and the widths of the triangles were 860 mels. On the opposite side of the series, for 26 filters the spacing was 80 mels and the widths of triangles were 160 mels. The results of all
experiments are given in Table 1. Since the best $Acc$ oscillates between 82 and 83.5%, we indicated in Table 1, for better survey, by a gray color the items with recognition accuracy higher than 82%. The dependency of the best $Acc$ for a given number of filters is depicted in Figure 1 (for the highest $Acc$ see the rows of Tab 1). The figures attached to the drawn curve indicate the numbers of coefficients for which these results were obtained.

**Figure 1. Dependency of the best $Acc$ to the number of filters used in the MFCC parameterization.**

Similarly, the dependency of the best results of $Acc$ for a given number of coefficients is shown in Figure 2 (for the highest $Acc$ see the columns of Tab 1). The figures attached to the curve indicate the number of corresponding filters. It is evident from Table 1, that the optimal setting of the front-end working with the MFCC parameterization can be reached with much less number of filters than it is recommended by the theory of the critical bandwidths. We can see two relatively stable areas of the best $Acc$ ($9 \times 12$ filters and $7 \times 3 \times 10 \times 3$ coefficients or $16 \times 24$ filters and $12 \times 3 \times 16 \times 3$ coefficients), where the first one does not quite satisfy the theory of critical bandwidths. A stable and very high $Acc$ was obtained, for example, already for 9 filters and $7 \times 3 = 21$ coefficients. Such a setting brings considerable computation savings achieved during parameter calculations and log-likelihood estimations.

**Figure 2. Dependency of the best $Acc$ on the number of coefficients used in the MFCC parameterization.**

### 4.2. Influence of the coefficient and filter numbers on recognition accuracy for the PLP parameterization

In this case the experiments were realized for increasing number of filters from 5 to 23 (with the step 2) and for increasing number of coefficients from $4 \times 3 = 12$ to $12 \times 3 = 36$. Conditions connected with these experiments were analogical as in case of the MFCC parameterization. Results achieved are given in Table 2. In Figures 3 and 4 similar dependencies as for the MFCC parameterization are drawn. From Table 2 and Figures 3 and 4 we can see only one very stable area of the high $Acc$ (better than 82%), contritely for $11 \times 23$ filters and $7 \times 3 \times 12 \times 3$ coefficients. It is evident that the PLP parameterization satisfies much better the critical bandwidths theory.

**Table 2. Recognition results for various numbers of filters and parameters in PLP-cepstral representation**

<table>
<thead>
<tr>
<th>Number of coefficients</th>
<th>12</th>
<th>15</th>
<th>18</th>
<th>21</th>
<th>24</th>
<th>30</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of filters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>62.4</td>
<td>62</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>76.3</td>
<td>78.7</td>
<td>75.8</td>
<td>75.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>77.0</td>
<td>81.3</td>
<td>81.8</td>
<td>80.9</td>
<td>79.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>75.7</td>
<td>79.5</td>
<td>82.4</td>
<td>82.6</td>
<td>83.2</td>
<td>82.3</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>75.4</td>
<td>79.8</td>
<td>81.8</td>
<td>82.6</td>
<td>83.1</td>
<td>83.7</td>
<td>82.2</td>
</tr>
<tr>
<td>15</td>
<td>74.9</td>
<td>80.4</td>
<td>81.9</td>
<td>83.5</td>
<td>83.3</td>
<td>83.3</td>
<td>82.2</td>
</tr>
<tr>
<td>17</td>
<td>74.6</td>
<td>80.9</td>
<td>80.9</td>
<td>83.0</td>
<td>82.9</td>
<td>82.6</td>
<td>82.0</td>
</tr>
<tr>
<td>19</td>
<td>75.2</td>
<td>80.9</td>
<td>82.2</td>
<td>83.5</td>
<td>82.9</td>
<td>83.5</td>
<td>82.6</td>
</tr>
<tr>
<td>21</td>
<td>74.3</td>
<td>80.5</td>
<td>81.6</td>
<td>82.9</td>
<td>83.9</td>
<td>83.1</td>
<td>83.0</td>
</tr>
<tr>
<td>23</td>
<td>75.2</td>
<td>80.6</td>
<td>81.8</td>
<td>82.5</td>
<td>83.5</td>
<td>83.2</td>
<td>83.3</td>
</tr>
</tbody>
</table>

**Figure 3. Dependency of the best $Acc$ to the number of filters used in the PLP parameterization.**

**Figure 4. Dependency of the best $Acc$ on the number of coefficients used in the PLP parameterization.**

### 4.3. Influence of acceleration coefficients and the mean or amplitude normalization of coefficients to the accuracy

In a following set of experiments there were tested an influence of acceleration coefficients and an effect of an amplitude (AmNo) and a cepstral mean (MeNo) normalization processes to the level of the $Acc$. All experiments were performed for small but effective sets of parameters derived in both parameterization techniques from corresponding 7 static coefficients ($C_7$). Let us mention that
CA14 means 7 static+7 delta coefficients, CAΔ23 indicates 7 static+7 delta+7 delta-delta coefficients. Both cepstral mean normalization (for MFCC and PLP cepstral representation) and amplitude normalization were performed with the CAΔ23 sets. From the recognition results given in Table 3 we can deduce the stable and evident benefit of especially additional delta coefficients (which help to increase the Acc by about 25% on average) while using additional delta-delta coefficients improved the accuracy only by 1 or 2% on average. The benefit of the mean cepstral normalization in improving the accuracy was usually higher than 2.0%.

### Table 3. Effect of further processing steps on Acc

<table>
<thead>
<tr>
<th>C7</th>
<th>CA14</th>
<th>CAΔ23</th>
<th>MeNo</th>
<th>AmNo</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>55.1</td>
<td>80.3</td>
<td>83.1</td>
<td>84.2</td>
</tr>
<tr>
<td>MFCC</td>
<td>55.2</td>
<td>81.0</td>
<td>82.2</td>
<td>84.4</td>
</tr>
<tr>
<td>MFCC</td>
<td>52.2</td>
<td>78.9</td>
<td>81.3</td>
<td>84.1</td>
</tr>
<tr>
<td>PLP</td>
<td>54.0</td>
<td>79.8</td>
<td>80.9</td>
<td>83.9</td>
</tr>
<tr>
<td>PLP</td>
<td>55.0</td>
<td>80.2</td>
<td>82.6</td>
<td>84.9</td>
</tr>
<tr>
<td>PLP</td>
<td>55.3</td>
<td>81.1</td>
<td>83.0</td>
<td>85.5</td>
</tr>
</tbody>
</table>

### 5. Conclusions

This paper discussed results of a large number of experiments with the telephone-based speaker independent continuous speech recognition. Two parameterization techniques – the MFCC and PLP – were tested and compared. The recognition results showed that both parameterizations are comparable but the PLP one provides slightly better and robust (stable for a larger number of parameters – the number of coefficients and filters) results. In future work we would like to utilize a profit of the delta-delta coefficients but with reduced number of components.

### 6. Acknowledgement

Support for this work was provided by the Ministry of Education of the Czech Republic, project no. MSM234200004, and by the Grant Agency of the Czech Republic, project no. 102/96/K087.

### 7. References


