Comparative Wavelet and MFCC Speech Recognition Experiments on the Slovenian and English SpeechDat2

Robert Modic*  Borge Lindberg#  Bojan Petek*

*Interactive Systems Laboratory  #Center for PersonKommunikation  University of Ljubljana, Slovenia
robert.modic@guest.arnes.si  bojan.petek@uni-lj.si  bli@cpk.auc.dk

1. Introduction

The main motivation for this project was to study performance of non-linear speech analysis methods in automatic speech recognition. Specifically, we selected wavelet transform as a promising non-linear tool for signal analysis that has been already successfully applied in many tasks, such as in image recognition and compression leading to standards such as JPEG2000. The plan was to perform a comparative analysis between the standard mel-cepstral and wavelet based set of features and to evaluate the baseline speech recognition rates of two aforementioned parameterization methods.

We start with a brief description of the Fourier and wavelet transforms from the perspective of joint time–frequency analysis where we focus on localization issues of the two transforms. Ability of the transformation to properly capture short time events is defined with the localization capabilities of its basic functions and is one of the prerequisites for a successful application in speech processing. The Fourier transform offers constant time–frequency resolution where the wavelet transform enables better frequency resolution at low frequencies and better time localization of the transient phenomena in the time domain [1]. This very much resembles to the first stage of human auditory perception [2] and to basilar membrane excitation [3] where the wavelet transform introduces roughly logarithmic frequency sensitivity. We carried out comparative within and cross-language experiments on the Slovenian and English SpeechDat2 [4] databases using the standard mel–cepstral and the wavelet based set of features. The tool used in automatic speech recognition was the reference recogniser [5,6] that is built around the HTK toolkit. This enabled us to conduct controlled experiments on six different subsets of SpeechDat2 vocabularies (yes/no sentences, citinames, phonetically rich word, digits, etc).

2. Wavelet Packet Parametrization

While the mel–cepstral parameterization of speech is an integral part of the HTK itself, we had to devise a suitable technique to implement the wavelet parameterization. We have chosen the MATLAB since it offers interactivity and supports the wavelet transformation computation within the Wavelet Toolbox. Additionally, we also used the Wavelab802 package [7]. Wavelet packet transform offers ability to arbitrary split the time–frequency axes. In order to achieve similar frequency decomposition as used in the mel-scale parametrization we used wavelet packet perceptual decomposition tree that was first proposed by R. Sarikaya [8] and yields the wavelet packet parameters (WPP).

The mother wavelet chosen in signal decomposition was the Daubechies compactly supported wavelet with two vanishing moments [9]. Daubechies wavelets are optimal in a sense that they offer minimal support of 2p for the given number of vanishing moments p. This also enabled a fast computation and decomposition using perfect reconstruction filterbank also called a conjugate mirror filter. We devised scripts and the MATLAB routines to fully embed the wavelet speech parameterization into the reference speech recognizer. Since the interpreter tool was found to be too slow given the size of the database we also had to find an appropriate solution to speed up the wavelet feature computation. This was achieved by using the MATLAB compiler. Since we were unable to compile the default Wavelet Toolbox function for the wavelet packet decomposition we resorted to use the Wavelab802 instead.

3. Experiments

Experiments included comparative evaluations of the recognition results using the mel–cepstral and the wavelet parametrizations on the 1000-speaker Slovenian and English SpeechDat2 databases. In order to analyze the baseline recognition performance that

---

1 Young Investigator supported by the MŠZŠ of the Republic of Slovenia and Socrates/Erasmus exchange student under the multilateral agreement UL D-IV-1/99-JM/Kc. Research was also supported in part by the COST 277 project.
would reflect the differences in frequency decomposition between the MFCC- and WPP parametrizations, it was decided that no dynamic information should be included into the feature vectors. It is well known that the delta mel–cepstral features improve the performance of hidden Markov models, yet in our experimental setup, we rather aimed to analyze how the inherent underlying transformation differences influence the MFCC and WPP-based recognition performance. That turned us away from using deltas.

Comparison of the recognition performance differences between Slovenian and English SD2 databases were aimed to provide information concerning the robustness of features to noise. We determined a significant difference in global average SNR on the Slovenian and English SD2, i.e., 25.8 dB and 40.1 dB SNR were estimated, respectively.

We used 25 ms speech window with mel–cepstral and 32 ms window with wavelet features, due to specific decomposition structure. Feature vectors were of length 24 for both parametrizations. We also used the same skip rate of speech window with the value of 10 ms. This ensured a fair comparison between the mel–cepstral and wavelet speech recognition experiments. Speech feature vector computation included calculation of log energies in the mel–scaled filterbanks. The mel–scaled distribution of wavelet bandpass filter was achieved using the perceptual wavelet decomposition structure [8]. Log filterbank energies and decorrelation with the DCT were used to produce mel–cepstral features. On the other hand, the wavelet transform was applied to decorrelate and yield the wavelet packet parameters (WPP).

4. Results

Results obtained in our experiments are shown in Figures 1 and 2 below.

![Figure 1](image1.png)

**Figure 1** Word error rates for six standard tests on the Slovenian SD2 using the mel–cepstral (MFCC) and wavelet (WPP) features.

![Figure 2](image2.png)

**Figure 2** Word error rates for six standard tests on the English SD2 using the mel–cepstral (MFCC) and wavelet (WPP) features.

5. Discussion

Tests that give the most relevant information are the tests on city names (citynames) and the phonetically rich words (rwords). These tests are representative due to the diverse phonetic content and can serve as baseline for judging the overall success of the parameterization methods involved. Slovenian SD2 experiments exhibit a small improvement of the recognition results with the wavelet features on citynames and rwords. We could hypothesize that the variable frequency resolution in the wavelet transform enhances the overall recognition rate.

We also tested the recognition performance using a 32 ms speech window in the MFCC calculation. Unequal MFCC and WPP window durations were therefore not considered to be problematic since the MFCC recognition scores were found to be consistently worse for longer window durations.

The English SD2 experiments yielded consistently better results obtained by the wavelet features. This could imply that the wavelet features are more robust in the variable noise conditions.

During the experiments we observed the appearance of side lobes in band pass filters that cut out frequency content of the signal. This is due to the non-optimal separability of conjugate mirror filter that implements the Daubechies 2 mother wavelet. Another observation was a different phone level alignment between MFCC and WPP features.

Additionaly we experienced a problem when we applied a threshold to small values of energies before the log followed by a de-correlation with the wavelet transform was to be taken. Log tends to boost small values. Since these values presumably belong to noise they represent the additional data that the model has to...
absorb. This possibly yielded to degraded overall recognition performance. The empirical threshold we've used with the Slovenian SD2 wouldn't work well for the English SD2. The HTK couldn't cope with the English WPP features calculated by thresholding and reported an "overprunning" error which was remedied by the removal of thresholding.

In conclusion, despite the preliminary stage of our experimental setup in the field of non-linear speech analysis, the results confirmed the hypothesis that using wavelets may bring potential in automatic speech recognition. Further work and improvements should incorporate the use of delta and delta–delta coefficients. The phoneme classification experiment within and between languages could also be considered in order to give additional information on the specific properties of parameterization techniques. Since SpeechDat2 represents a noisy telephone database the use of wavelet de-noising could offer a solid foundation to increase the robustness of wavelet parameterization method to noise and additionally improve the recognition results.

6. References


