A Comparison of Several Approaches to the Feature Extractor Design for ASR Tasks in Telephone Environment

A. Gallardo-Antolín[†‡], J. Macías-Guarasa[‡], J. Ferreiros[‡], R. Córdoba[‡], J. M. Montero-Martínez[‡], R. San-Segundo[‡] and J. M. Pardo[‡]

† Departamento de Teoría de la Señal y las Comunicaciones, EPS, Universidad Carlos III de Madrid, Spain

‡ Grupo de Tecnología del Habla. Dpto. Ingeniería Electrónica, ETSIT, Universidad Politécnica de Madrid, Spain

E-mail: gallardo@tsc.uc3m.es, {macias, jfl, cordoba, juancho, lapiz, pardo}@die.upm.es

ABSTRACT

Automatic speech recognition (ASR) systems are usually composed of a parameterization module and a back-end classifier. The performance of the overall system strongly depends on the choice of the feature extraction module. In this paper we investigate two different approaches for designing this module. In the first one (the conventional approach), its main characteristics are chosen based on psychoacoustic knowledge. In the second one, a data-driven technique ("Discriminative Feature Extraction"-DFE-), which performs a simultaneous optimization of the feature extractor and the back-classifier, is used. Both strategies have been applied to a front-end based on the Wavelet Transform (WT). Results show that DFE systematically improves the performance. In fact, applying the DFE strategy to the WT-based acoustic features, a relative error reduction around 23% (compared to the conventional features based on Short-Time Fourier Transform) is achieved when using the SpeechDat database with a vocabulary of 1000 words.

1. INTRODUCTION

The performance of the ASR systems strongly depends on the choice of the feature extraction module. This module can be designed using two different strategies. The first one (the conventional approach) is based on heuristics or psychoacoustic knowledge. In the second one, some of the characteristics of the front-end are training according to a certain minimization criterion. The Discriminative Feature Extraction (DFE) method [1] provides an appropriate formalism for this strategy.

In this paper we explore the application of DFE for optimizing a wavelet-based front-end for an ASR system working in a telephone environment.

In most of the current ASR systems, feature extraction techniques are based in the analysis of the speech waveform on short time windows (typically, 20-30 ms) using the Short-Time Fourier Transform (STFT). However, several studies show that the information obtained on different time scales or resolution levels could improve the recognition accuracy.

The Wavelet Transform (WT) offers an implicit way to exploit information on multiple time scales or resolutions. The WT has been used instead of the DCT involved in cepstral computation [2] or the conventional STFT [2], [3], [4]. Our wavelet-based front-end is related to this latter approach.

The paper is organized as follows. Section 2 presents the wavelet-based parameters (MWCC) in comparison to the conventional MFCC. Section 3 reviews the DFE formalism and the application to the wavelet-based system. Section 4 presents the experiments and discusses the results. Finally, some conclusions are drawn in section 5.

2. PARAMETERIZATIONS BASED ON PSYCHOACOUSTIC KNOWLEDGE

Usually, the design of the front-end extractor is based on psychoacoustic knowledge or heuristics. An example is the number of critical bands and its values for the extraction of MFCC parameters or the number of critical bands an its values for MWCC parameters (see below).

Along this section, we describe the features investigated in this paper: cepstral parameters based on STFT (MFCC) and cepstral parameters based on WT (MWCC).

2.1. Mel-Frequency Cepstral Parameters (MFCC)

There are two main stages in MFCC-based parameterization. The first step is the calculation of the log filterbank energies of the signal. They are derived using the STFT defined by,

$$S_x(\tau, f) = \int_{-\infty}^{\infty} x(t+\tau)\, h(t)\, e^{-j\,2\pi f t}\, dt \qquad (1)$$

where $x(t)$ is the speech signal and $h(t)$ represents a window function. Thus, a version of the signal windowed by $h(t)$ around time τ is analyzed at all frequencies f.

ISBN 1-876346-48-5

Once a window has been chosen for the STFT, then the time-frequency resolution is fixed over the entire time-frequency plane since the same window is used at all frequencies. This is one of the main problems associated with this approach.

The log filterbank energies are obtained passing the power spectrum, $|S_x(\tau, f)|^2$ through a mel-scaled filterbank and using a log function.

The second stage consists of the decorrelation of the log-energies using a DCT to obtain MFCCs. Finally, the first derivatives of MFCCs (ΔMWCC) are appended to the feature vector.

2.2. Wavelet-based Cepstral Parameters (MWCC)

The extraction of MWCC parameters also involves the two main steps described above. The main difference is that the log filterbank energies are obtained using the Wavelet Transform (WT) instead of the STFT.

The Continuous WT of a signal *x(t)* is defined as,

$$W_x(\tau, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\, \Psi^*_{a,\tau}(t)\, dt \qquad (2)$$

where $\Psi_{a,\tau}(t)$ are the scaled (by a) and translated (by τ) versions of the basic wavelet $\Psi(t)$. In our case, it is the Morlet wavelet (a modulated Gaussian function). So, equation (2) can be expressed as,

$$W_x(\tau, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t+\tau)\, g\left(\frac{t}{a}\right) e^{-j 2\pi \frac{f_o}{a} t}\, dt \qquad (3)$$

in which *g(t)* is a Gaussian window and f_o = 3 Hz. Although *g(t)* is an infinite-length function, only the values above 0.1 are kept, so the size of Ψ (t) is $\lambda_0 \approx 4.625$ s.

Comparing (1) and (3), it can be observed that both expressions are similar with $f = f_o / a$ and $h(t) = g(t/a)$ [4]. In fact, in the case of WT, at certain scale *s*, the signal is windowed with a Gaussian function, $g(t/a)$ (with length $\lambda_s = a.\lambda_0$), and then analyzed at the frequency $f_s = f_o / a$.

The main difference with the classical STFT is the analysis window length: constant for all frequencies in STFT and variable with the scale factor *a* (and hence, with frequency) in WT, thus enabling different time/frequency resolutions. For small values of *a*, WT analysis provides a good temporal resolution for high frequencies (λ_s decreases and f_s increases) and, for large values of *a*, a good frequency resolution for low frequencies is obtained (λ_s increases and f_s decreases).

In our case, we have used N_s = 34 scales corresponding to a set of 34 frequencies, f_s, emulating the mel scale. For each scale, $a = f_o / f_s$, the size of the analysis window was calculated according to $\lambda_s = a.\lambda_0$ (where λ_0 is the size of the basic wavelet).

As in MFCCs, the first derivatives of MWCCs are computed and appended to the feature vector (ΔMWCC).

2.3. Combination of acoustic features

The combination of different types of features provides error rate reductions if they have significant complementary information contents. For investigating this possibility, a comparative analysis of the errors made by MFCC, ΔMFCC, MWCC and ΔMWCC-based systems were conducted. These preliminary results showed that MWCC and ΔMFCC are the best candidates for merging: In our experimentation (see section 4.1) 31.5 % of the errors made by the MWCC-based system would be corrected by ΔMFCC-based one while 26.8% of ΔMFCC-based system are correct decisions in the MWCC-based one. It will be shown in section 4.2 that the strategy of feature combination outperforms the baseline system.

2.4. Temporal filtering of time trajectories

Both, MFCC and MWCC parameters, can be improved by applying a temporal filtering of their time trajectories. In particular, we have used the CMN technique ("Cepstral Mean Normalization").

3. PARAMETERIZATIONS BASED ON DISCRIMINATIVE FEATURE EXTRACTION

Usually, in ASR systems, the feature extraction and the back-end classifier modules are designed independently. However, this procedure doesn't guarantee an error-rate reduction of the overall system. One of the methods that performs simultaneously the optimization of both modules is the Discriminative Feature Extraction technique (DFE).

Recently, it has been applied to obtain optimal filter-banks [1] and lineal transformations of the feature space [5]. In this paper, we propose the DFE-method for training the wavelet-based front-end described in section 2.2.

3.1. Adaptive Gaussian wavelets

The Adaptive Gaussian wavelets can be considered a generalization of the Morlet wavelet, in which the modulation frequency (f_o) is a variable value that can be adjusted according to a certain minimization criterion [6]. As a consequence, for a certain scale the value of the window size is independent of the frequency. The Adaptive Gaussian Wavelet Transform is defined as,

$$W_x(\tau, s) = \int_{-\infty}^{\infty} x(t+\tau)\, w_{g,s}(t)\, e^{-j 2\pi f_s t}\, dt \qquad (4)$$

in which $w_{g,s}(t)$ is an adaptive Gaussian window with length λ_s modulated at the frequency f_s.

3.2. Review of the DFE algorithm

DFE is usually performed by using the MCE-GPD algorithm ("Mininum Classification Error/Generalized Probabilistic Descent") [7]. According to this approach, the set of trainable parameters, Φ, are iteratively re-estimated in order to minimize a certain average loss function, $L(\Lambda)$, which is a good approximation of the classification error rate of the training data.

For doing it, the GPD algorithm is applied. In this way, Φ is iteratively updated along a gradient descent direction. At iteration k, the new set of parameters is calculated as,

$$\Phi^k = \Phi^{k-1} - \eta \nabla L\left(\Phi^{k-1}\right) \qquad (5)$$

where η is the learning step size and controls the convergence of the algorithm. The gradient $\nabla L(\Phi)$ is obtained by computing the partial derivatives of $L(\Phi)$ via the chain rule.

For the wavelet-based front-end, the trainable parameters, Φ, to be adjusted by DFE are $\Lambda = \{\lambda_1,..., \lambda_{Ns}\}$, where λ_s represents the window size for the frequency f_s of the corresponding Adaptive Gaussian wavelet.

The DFE algorithm needs a set of labeled sequences of training vectors, each of them belonging to a predefined class. In our DFE implementation, the classes are defined by HMM states. So, initially, the training data must be segmented and labeled into states (in our case, by using the Viterbi algorithm). Obviously, as the back-end HMM classifier is not very optimized at this stage, the initial segmentation contains incorrect labels that can affect the behaviour of DFE. Before performing DFE, the updated HMM recognizer segments the training sequences in a more accurate way and this contributes to improve the performance of DFE in the next iteration. This practical implementation of DFE is called "Segmental DFE" [5]. In summary, in order to reduce the error rate of the overall ASR system, it is necessary to iteratively segment the training data, adjust the parameters of the feature extractor module using DFE and adjust the parameters of the HMM classifier using the ML procedure ("segmental iteration"). The global process ends when the reduction of the average loss function is smaller than a predefined threshold.

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Database and baseline system

In our experimentation, we have used part of the SpeechDat database [8], a speaker-independent speech corpus collected over the Spanish telephone network. It contains utterances pronounced by 1000 different speakers and it has been recorded at 8 KHz (A law). The training and test sets consist of 5080 and 2003 utterances, respectively The dictionary is composed by 1000 words.

The baseline is a speaker-independent, isolated-word HMM-based ASR system. Context-independent CDHMM with three states and one Gaussian mixture per state are used for modeling each of the 47 allophone-like units. So, the total number of states (different classes in DFE) is 141.

In order to state the statistical significance of the experimental results shown in the following subsections, we have calculated the 95 % confidence intervals.

4.2. MFCC vs. MWCC

For the baseline system, the speech signal is analyzed with a 25 ms Hamming window every 10 ms. For the computation of the log-energies, 17 triangular mel-spaced filters were used. Finally, 10 MFCCs and the frame log-energy and their first derivatives are computed.

For the wavelet-based system, the parameters were also extracted every 10 ms. We used 34 scales, so the speech signal is analyzed with 34 analysis windows of variable size (from 50 to 4 ms) as represented in Figure 1 (labeled as "Initial MWCC"). Preliminary results suggested to constrain the maximum window length to 50 ms. Finally, 10 MWCCs and the log-energy (extracted from a window of 25 ms length) and their derivatives are computed.

In Table 1, we compare the recognition rates for the MFCC and MWCC-based systems. MWCC+ΔMWCC parameters outperform the baseline system, although the differences are not statistically significant.

Features	Recognition Rate (%)
MFCC + ΔMFCC (Baseline)	73.20 % [71.35% - 75.05%]
MWCC + ΔMWCC	75.48 % [73.68% - 77.28%]
MWCC + ΔMFCC	77.12 % [75.37% - 78.87%]

Table 1: Comparison of recognition rates for different parameterizations.

Based on the preliminary experimentation in section 2.3 we decided to combine MWCC and ΔMFCC features into one single stream. From Table 1 it is seen that the combination ("MWCC+ΔMFCC") results in improved performance compared to both, the baseline and MWCC-based systems.

4.3. Experiments with DFE

Since DFE uses a gradient descent search, which only guarantees to find a local minimum, a good initialization is recommended. In our case, the initial values for Λ were the indicated in section 4.2. According to a series of preliminary experiments we fixed the slope of the sigmoid $\alpha = 0.1$ and the learning rate $\eta = 0.0001$.

ISBN 1-876346-48-5

We have carried out two different experiments. In the first one, DFE was performed using the parameterization MWCC-ΔMWCC. In the second one, DFE used the combination of MWCC-ΔMFCC. In the first case, the algorithm reached the convergence after seven segmental iterations. In the second one, 20 iterations were needed for getting a negligible reduction of the average cost.

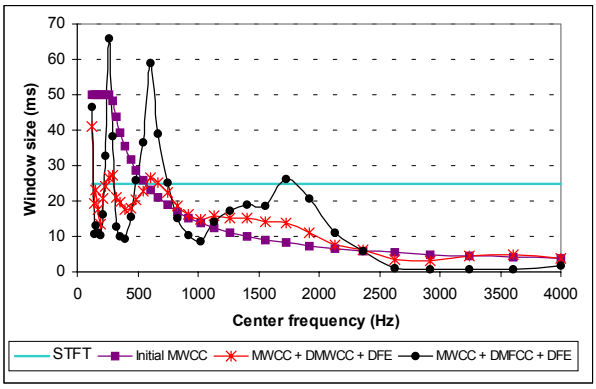

Figure 2: Window sizes vs. frequency: STFT, initial MWCC, MWCC + Δ MWCC (at segmental iter. 7) and MWCC +ΔMFCC (at segmental iter. 20).

Figure 2 shows the window sizes corresponding to each frequency band for both experiments. For comparative purposes, this figure also shows the sizes for the STFT-based system and the initial configuration of the MWCC-based system. It can be observed that windows corresponding to frequencies around 250, 650 and 1750 Hz become larger when applying the DFE procedure. On the other hand, windows corresponding to frequencies above 2250 Hz become shorter.

	Recognition Rate (%)	
Features	No CMN	CMN
MFCC + ΔMFCC (Baseline)	73.20 % [71.35% - 75.05%]	78,30 % [76,58% - 80,02%]
MWCC + ΔMWCC + DFE	76.98 % [75.22% - 78.74%]	79,89% [78,22% - 81,56%]
MWCC + ΔMFCC + DFE	79.42 % [77.73% - 81.11%]	81,78 % [80,17% - 83,39%]

Table 2: Comparison of recognition rates with different parameterizations and DFE.

Table 2 shows the recognition rates. As it can be observed in Table 2 (column labeled as "No CMN"), DFE (without and with feature combination) systematically improves the performance of the MFCC and the original MWCC-based systems (see Table 1). Again, best results are achieved using the feature combination strategy (plus DFE). In this case, relative error reductions around 23% compared to the baseline system are obtained and the differences in ASR performance are statistically significant.

Also in Table 2 (column labeled as "CMN") we show the results obtained when applying CMN to the acoustic features. It can be observed that CMN increases the recognition rate in all cases. Again, best results are obtained when applying the DFE method.

5. CONCLUSIONS AND FURTHER WORK

In this paper, we have compared the performance of a cepstral parameterization derived from the Wavelet Transform and the conventional MFCC parameters. We have shown that MWCCs show a similar performance compared to MFCCs in a telephone environment. However, the combination of both kinds of parameters (in particular, MWCC + ΔMFCC) outperforms the baseline system. We have also designed a procedure based on the DFE algorithm for training the optimal window sizes in the MWCC-based systems. In this case, the proposed method provides significant improvements in the ASR performance.

We are presently exploring the performance of the wavelet-based features in additive noise conditions.

6. REFERENCES

[1] Biem A., Katagiri S., McDermott E. and Juang, B.-H., "An Application of Discriminative Feature Extraction to Filter-Bank-Based Speech Recognition', *IEEE Trans. Speech and Audio Proc*., vol. 9, no. 2, 96-110, 2001.

[2] Sarikaya R., Pellom, B. L. and Hansen, J- H. L., "Wavelet Packet Transform Features with Application to Speaker Identification", *EUROSPEECH-01*, 2001.

[3] Gemello, R., Albesano, D., Moisa, L. and De Mori, R., "Integration of Fixed and Multiple Resolution Analysis in a Speech Recognition System", *ICASSP-01*, 2001.

[4] Wassner H. and Chollet, G., "New Cepstral Representation using Wavelet Analysis and Spectral Transformations for Robust Speech Recognition", *Proc. of ICSLP-96*, 1996.

[5] de la Torre, A., Peinado, A. M., Rubio, A. J. and Segura, J. C., "Discriminative Improvement of the Representation Space for Continuous Speech Recognition", in *"Computational Models of Speech Pattern Processing"*. NATO ASI Series, Springer Verlag, 1999.

[6] Kadambe S. and Srinivasan P., "Applications of Adaptive Wavelets for Speech", *Journal of Optical Engineering. Special Issue on Wavelets*, July 1994.

[7] Juang B. -H. and Katagiri S., "Discriminative Learning for Minimum Error Classification", *IEEE Trans. on Signal Processing*, vol. 40, nº 12, pp. 3043-3054, 1992.

[8] Moreno, A. "SpeechDat Documentation [cd-rom]", ver 1.0, Universitat Politècnica de Catalunya, 1997.