NEW CEPSTRAL REPRESENTATION USING WAVELET ANALYSIS AND SPECTRAL TRANSFORMATION FOR ROBUST SPEECH RECOGNITION

Hubert Wassner *, Gérard Chollet *,**.

* IDIAP, CP 592, 1920 Martigny, Switzerland
** CNRS URA-820, ENST, Paris, France
e-mail: wassner@idiap.ch, chollet@idiap.ch, chollet@sig.enst.fr

ABSTRACT

The goal is to improve recognition rate by optimisation of Mel Frequency Cepstral Coefficients (MFCCs): modifications concern the time-frequency representations used to estimate these coefficients. There are many ways to obtain a spectrum out of a signal which differ in the method itself (Fourier, Wavelets,...), and in the normalisation. We show here that we can obtain noise resistant cepstral coefficients, for speaker independent connected word recognition.The recognition system is based on a continuous whole word hidden Markov model. An error reduction rate of approximately 50% is achieved with word models.

1. INTRODUCTION

The subject is about optimizing cepstral estimation for speaker independent continuous speech recognition using HMMs. These adaptations take place in the first stage of cepstral calculation, the time frequency transformation.

This paper points out that gains can be obtained by choosing the time-frequency transformation and its normalisation. We study the most often used coefficients: the MFCCs (Mel Frequency Cepstral Coefficients). The first part of this paper is a short reminder of the classical computation method for these coefficients. The second part is the description of the different improvements proposed here. The last part exposes the database used for the tests and the results.

2. MFCC estimation

MFCCs are used to describe the short-term spectral envelope of a speech signal. Several studies have shown the importance of using a Mel frequency scale. There are two main steps in calculating MFCCs:

- Calculating the log-magnitude spectrum out of a filter-bank.

  This step can be simulated by computing the power spectrum, passing it through a filter-bank and using a log function.

- Calculating the cosine transform of the filter-bank output.

The figure 1 presents the different stages of Cepstrum computation. For more information see [6].

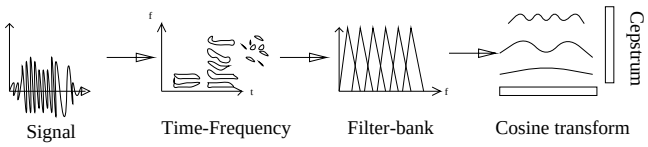

Figure 1: Signal to cepstrum

3. Possible Improvements

The power spectrum is often estimated by FFT, but this may not be the best time-frequency transformation. For instance a wavelet transform can be used to obtain the spectrum, with a different time-frequency accuracy compromise. We use here a wavelet transform defined by M Unser [7]:

$$W_x(t,a) = \frac{1}{\sqrt{a}} \sum_{T=-\infty}^{+\infty} x(t+T) g(a,T) e^{\frac{-i\Omega T}{a}}$$

In this formula : a is the scale factor, linked to the frequency by :

$$f = \frac{\Omega}{2\pi a}$$

g is a window the size of which depends on a, according to :

$$g(a,t) = \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2a^2}}$$

A good choice of the scale factor a allows to simulate a Mel scale filter bank.

We can compare this transfom with the short term Fourier transform [4] defined by:

$$S_x(t,f) = \sum_{s=-\infty}^{+\infty} h(s) x(t+s) e^{-2i\pi f s}$$

(where h(s) is a window, for instance a gaussian). The latter can be seen as a Wavelet transform, with

$$a = \frac{\Omega}{2\pi f}$$

and g(a,f) = h(t).

The main difference is the size of the time window: constant for Fourier and variable for wavelets. Likewise the notion of scale in wavelets can be seen as a change of variable in Fourier analysis.

On the two figures 2, 3 we can notice the different time/frequency accuracy compromises between the two methods.

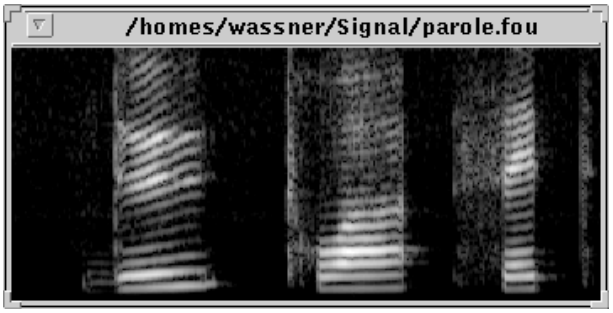

Figure 2: Fourier analysis of speech.

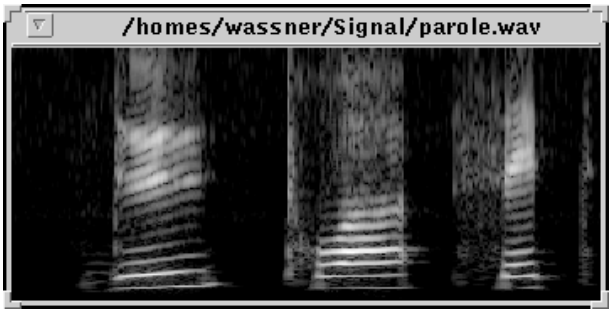

Figure 3: Wavelet analysis of speech.

In the classical cepstrum calculation, a log transformation is used to modify the power spectrum. Generally this transformation is done after the filter-bank. We will see that performing this transformation before the filter bank is more interesting in our case. (See figures 4 and 5 to compare the two different filtered spectra)

Moreover we can notice that in certain noisy conditions this log-transformation has much too low energy dynamics (certain low energy time-frequency zones can be interpreted as noise). Therefore other energy-transformation functions (see figure 7) have been experimented:

$$log2(x) = (\frac{log(x)}{log(M)})^2 * log(M)$$

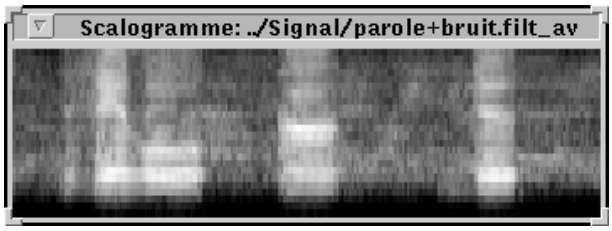

Figure 4: Filtered spectrum : log before filters.

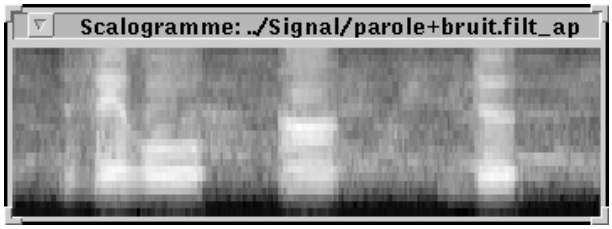

Figure 5: Filtered spectrum : log after filters.

$$sigm(x) = \frac{1}{1 + 0.0004 * e^{x/M * \alpha + 5}} * log(M)$$

Where M is the value of the maximum energy found in the time-frequency plane, from the speech signal studied.

The Sigmoid function allows to get rid of low energy and then to do an easy denoising transformation on the spectrum. This denoising seems very good on the picture of the time-frequency plane, see figure 6.

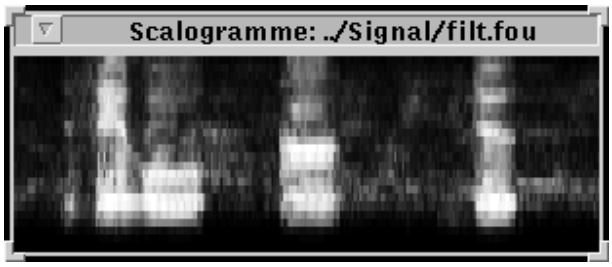

Figure 6: Filtered Spectrum with a sgimoid instead of a log, before the filter bank).

4. Experiments and results

4.1. Test Protocol

The test protocol was the same for all experiments, only the speech parametrisation was different.

The default HTK-like parameters used are:

- Mel scale
- Number of cepstrum coefficients : 12
- Number of filters : 24

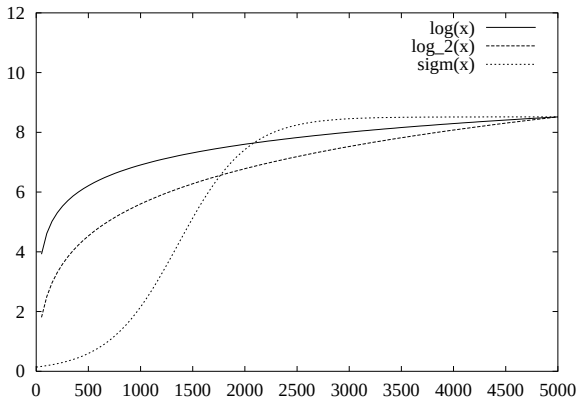

Figure 7: Different energy transformation functions.

- Cepstral liftering (sin) : 22
- Pre-emphasis coefficient : 0.98
- Frame-rate : 10 ms
- Window size : 25 ms for Fourier (variable for wavelets).

Each MFCC vector is composed of 12 cepstral coefficients, with energy, slope, and acceleration. This leads to a 39 elements vector per frame.

These parameters come from the HTK guideline [5]. Nothing proves that they are the best, but they are quite widely used.

Two of these experiments have been done using HTK's HCode program for parametrisation, in order to obtain references. The other experiments have been done using a specific program (HCepstre). The Markov model used are "left-right" models representing words. An initialisation and a re-estimation have been made on the models (HInit and HRest HTK(1.5) programs).

4.2. Database

Three sub-databases have been used, two extracted from "Polyphone" [2] and one from "Computer95". These two databases have been collected by IDIAP and the Swiss Telecom PTT from French spoken telephone speech. The words used are the french digits (0-9) for all the three sub-databases, added with the word "diese" (hash) and "etoile" (star) in the extracts from "Polyphone".The Polyphone database is low noise (people are calling from home). Computer95 is recorded from the annual computer forum at Lausanne, with a high background speech noise. The three sub-databases have different speakers (for more information see the table 1).

4.3. Results

Table 2 presents the main interesting experiments and their results (in percentage). First we can notice the important

	speaker	words
learning (Polyphone)	498	2962
Evaluation (Polyphone)	429	2574
Test (Computer95)	376	3760

Table 1: Databases' composition

gain compared to the reference experiment (1). The improvement seems larger on the Computer95 database. In fact the gain is a reduction of 50% of the number of errors on both Polyphone and Computer95.

Parametrisation		polyphone	Computer
1	MFCC reference (HTK)	92.70	65.48
2	MFCC log before filters	97.40	79.36
3	MFCC log after filters	91.18	62.13
4	MFCC sigmoid (60)	93.28	67.31
5	MFCC sigmoid (30)	96.70	75.11
6	MFCC sigmoid (15)	96.27	80.88
7	MFCC sigmoid (10)	95.18	81.33
8	MWCC, $\Omega = 9$, log	96.89	80.40
9	MWCC, $\Omega = 11$, log	97.47	82.61
10	MWCC, $\Omega = 9$, log2	96.70	81.62
11	MFCC, log2	93.59	71.86
12	MFCC (HTK) with small learning database	87.61	59.81
13	MWCC, $\Omega = 11$ with small learning database	96.15	80.16

Table 2: Results

Remarks:
- MWCC are mfcc calculated on a spectrum obtained by wavelet-transform.
- log, log2, sigmoid (with a parameter) are energy transformation functions.(this transformation is done before the filter-bank, except for the experiment 1,3,12.)
- Experiments 12 and 13 have been made using only 45% of the learning data base, in order to study the learning speed (in terms of database size)

The 95% confidence intervals [1] are [91.63, 93.64] on Polyphone, and [63.92, 67.00] on Computer95, for the first experiment and respectively [96.78, 98.00] and [81.34, 83.80] for the ninth. This may give an idea of the signifiance of the results.

We have noticed that putting the log function before the filter-banks leads to better results. Then we have choosen, for all the other experiments, to place the energy transformation before the filter-bank (except for the experiment 1,3,12). More information on this choice can be found in [8].

In experiments 4, 5, 6 , 7, where a sigmoid is used instead of a log, we can notice that for the sigmoid parameter α, the best value seems to be the same on the different databases. This may indicate the existence of an optimal function for all conditions.

The problem of the sigmoid function is that there is a compromise to be found:

- Low value of its parameter will highly denoise the picture of the time-frequency plane, and also seems to minimize the effect of noice on the Computer95 database.
- Too low values of its parameter distort too much the spectrum (see figure 8).

This compromise state explain why there is interesting gain on the Computer95 database in lowering the sigmoid parameter, but simultaneously results on the Polyphone database are getting down when this parameter passes a certain limit value.

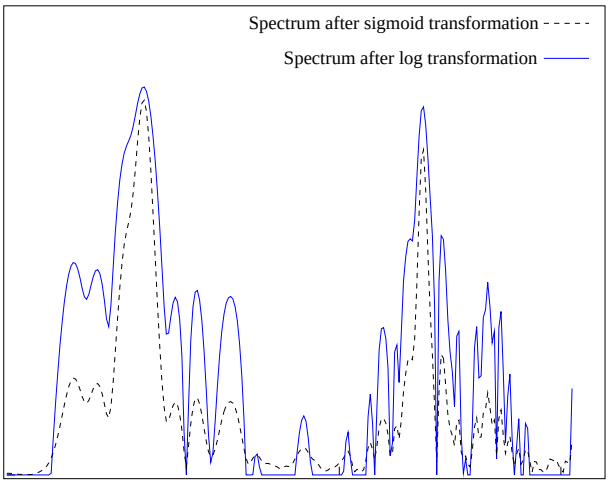

Figure 8: Spectrum deformation when using a sigmoid function instead of a log function.

Experiments 12 and 13 indicate the possibility of learning twice quicker, in terms of database size, with very little loss. Even with a halved database, we reach better results than with the classical MFCCs on the whole database.

We may also point out that the energy normalisation function is closely linked with the time-frequency method used to obtain the spectrum. There is a difference of 3% on Polyphone and 10% on Computer95 between the couples (MWCC, log2) and (MFCC, log2).

A context re-estimation (using HTK's HERest) with 390 new speakers (2340 words) taken from Polyphone database leads to the results reported in table 3.

Parametrisation		Polyphone	Computer95
1	MFCC reference (HTK)	96.81	76.54
9	MWCC, $\Omega = 11$, log	97.98	83.48

Table 3: Results after context re-estimation

The 95% confidence intervals are given in table 4.

	Polyphone	Computer95
Exp 1	[96.05, 97.42]	[75.14, 77.88]
EXp 9	[97.36, 98.45]	[82.24, 84.64]

Table 4: 95% Confidence intervals after context re-estimation.

The new parametrisation is again better,but the difference becomes very short on the Polyphone database. The difference is always important on the Computer95 database.

5. Conclusion

The cepstrum computation in its widely used form, appears clearly not to be an optimal solution. By keeping the same theoretical framework and calculating coefficients with more care, the cepstrum may give better results for recognition rate, specialy in noisy conditions. One improvement may be the research of the best couple time-frequency method, and normalisation function.

Acknowledgments

Thanks to Dominique Genoud and Robert Van Kommer for providing the Computer95 database and Michel Combot for a carefull review of early draft.

6. REFERENCES

1. G Chollet. New advances and trends in speech recognition and coding., chapter Evaluation of ASR systems, algorithms and databases. Nato ASI, Bubion, 1994.
2. G Chollet, JL Cochard, A Constantinescu, P Langlais, and R Van Kommer. Swiss french polyphone and polyvar : Telephone speech databases to model inter- and intra-speaker variability. Technical report, IDIAP, 1996.
3. Charles K. Chui. Wavelet Analysis and its Applications, vol 1: introduction to Wavelets. Academic Press, 1992.
4. Leon Cohen. Time-frequency distribution -a review. In Proc. IEEE, volume 77, july 1989.
5. Entropic Research Laboratory. Using HTK to design a Speaker independent connected digit recognition system, 1994.
6. L Rabiner and Biing-Hwang Juang. Fundamentals of Speech Recognition. Prentice-Hall signal processing series, 1993.
7. Michael Unser. Fast gabor-like windowed fourier and continuous wavelets transforms. IEEE Signal Processing Letters, 1(5), may 1994.
8. H Wassner. Etude sur la parametrisation du signal en reconnaissance automatique de la parole. Technical report, IDIAP, 1995.
9. H Wassner and G Chollet. Une nouvelle estimation de coefficients cepstraux pour la reconnaissance automatique de la parole. In JEP, Journees d'etude de la parole., 1996.