SPEECH ANALYSIS BASED ON MALVAR WAVELET TRANSFORM.

Christophe RIS, Vincent FONTAINE, Henri LEICH
Faculté Polytechnique de Mons, TCTS Labs,
Boulevard Dolez 31, B-7000 MONS, BELGIUM
e-mail : tcts@pip.fpms.ac.be

ABSTRACT

This paper presents a new pre-processing method developed with the objective to represent relevant information of a signal with a minimum number of parameters. The originality of this work is to propose a new efficient pre-processing algorithm producing acoustical vectors at a variable frame rate. The length of the speech frames is no longer fixed a priori to a constant value but results from a study of the signal stationarity. Both segmentation and signal analysis are based on Malvar wavelets since the orthogonal properties of this transform are the key to the problem of comparing measures done on frames of different lengths.

1. INTRODUCTION

The objective we followed in this work was to represent pertinent information of a signal with a minimum number of parameters. The most efficient way to avoid redundancy in signal representation is to perform the feature extraction on frames of signals that may have a variable length. To obtain such a pre-processing method, we proceed in two steps : the signal is first analysed on a set of pre-defined fixed length segments. Then some of these segments are merged to form an optimal segmentation in the sense of the optimisation of an entropy criterion.

Signal analysis over variable frame lengths is performed using the Malvar wavelets since the orthogonal properties of this transform cope with the problem of measurement done on variable length frames.

This paper is subdivided into six sections. After this short introduction, the second section introduces the Malvar wavelets and their advantages against other spectral transforms. The third section is divided into three parts. First we defines a new cepstrum referred to as Malvar cepstrum, derived from section two and used as feature vector in the recognition tasks. Second, we describe the segmentation algorithm and some expected advantages of speech recognition on variable length time-partition. And third, we depict possible approaches for the feature extraction. The fourth section presents some results of speech recognition based on this pre-processing. Finally, we give some conclusions.

2. BASIC PRINCIPLES

The Discrete Fourier Transform is well suited for spectral analysis of stationary signals since they can be decomposed in a canonical way in a linear combination of waves. For non-stationary signals, the DFT is not adapted at all because it gives only information on the spectral components appearing in the signal without any information on their temporal localisation. To avoid this weakness, DFT has been adapted for quasi-stationary signals (i.e. statistical properties are varying slowly with regard to observation period) to have some temporal information on spectral components. This gave birth to the Short Time Fourier Transform (STFT). In the STFT, the basis functions of the DFT are passed through a window $\omega(t)$ so that the basis functions are localised in time. It is now possible to determine when a particular frequency is present in the signal. On the other hand, frequencies composing the signal can not be determined as precisely as in the DFT (incertitude principle of Heisenberg). However, STFT constitutes a convenient tool to observe signal spectrums along the time axis.

In a general way, the time-frequency wavelet transform consists in decomposing the signal $s(t)$ into a linear combination of time-frequency units $u_R(t)$. This time-frequency units are defined on confined areas in the time-frequency plane. If we define such an area as $R=[a,b]x[\alpha,\beta]$, the function $u_R(t)$ must be contained in the interval $[a,b]$ for the time axis and in the interval $[\alpha,\beta]$ for the frequency axis. The decomposition can be written as :

$$s(t) = \sum_{k=0}^{\infty} c_k \cdot u_{R_k}(t) \quad (1)$$

A suitable approach for the wavelet decomposition consists in creating an optimised time-partition of the signal followed by a standard trigonometric transform (Fourier, DCT, DST, ...) :

- The signal is split in segments using a shifted window $\omega(t)$:

$$\omega(t-bl)\cdot s(t) \quad l = 0,\pm1,\pm2,\ldots \quad (2)$$

where b is the length of a segment

- The trigonometric transform (e.g. Fourier) gives:

$$S_l(k) = \int e^{-iakt}\cdot\omega(t-bl)\cdot s(t)\cdot dt \quad (3)$$

This expression is equivalent to the inner product of the signal$s(t)$ and the so-called wavelets:

$$u_{k,l} = e^{-iakt}\cdot\omega(t-bl) \quad (4)$$

This is the basic form of time-frequency wavelets. But this form leads to strong algorithmic difficulties and F. Low and R. Balian have proved [1] that, if $\omega(t)$ is sufficiently regular and localised in time, the set of functions $u_{k,l}$ will never provide an orthonormal basis of $L^2(\Re)$. This means that it is not possible to find a set of STFT basis functions $u_{k,l}(t)$ along the time axis that verifies equations (1) and (4). Only the rectangular window without overlapping regions eludes this demonstration.

A necessary condition to design orthonormal basis verifying equation (1) is that the basis functions of one frame are orthogonal not only to each other, but also to the functions in the surrounding frames. There exist many sets of functions verifying (1) among these : the system of Daubechies, Jaffard, Journé [1,2]. These transforms consists in applying a DCT and a DST alternatively coupled with an exponentially decreasing window. So doing, the segmentation using very regular windows can lead to an orthonormal basis of wavelets.

0-7803-2431-5/95 $4.00 © 1995 IEEE

Independently, Malvar developed a wavelet transform that also avoids the problems encountered with STFT. The Malvar Wavelet Transform consists in a non-redundant decomposition of any signal (of any duration) in a linear combination of elementary functions localised in time and frequency (eq. 1). The growing popularity of the Malvar system lies in the efficient algorithm developed in 1990 for computing the coefficients of the decomposition [4, 5].

The definition of this transform is based on some very simple and flexible constraints imposed in the choice of the window :

$$\begin{aligned}&\omega(t) = 0 \quad t \le -\pi \ or\ t \ge 3\pi \\ &0 \le \omega(t) \le 1 \text{ and } \omega(2\pi - t) = \omega(t) \qquad (5) \\ &\omega^2(t) + \omega^2(-t) = 1 \text{ if } -\pi \le t \le \pi\end{aligned}$$

An interesting advantage of these expressions is that they are very easily generalised to varying length windows. On an arbitrary time-partition $[a_j, a_{j+1}]$, we define the coefficients α_j and $\alpha_{j+1}>0$ so that $l_j=a_{j+1}-a_j>\alpha_j+\alpha_{j+1}$. The role of the coefficients is to allow the overlap of successive windows. The constraints proposed earlier has been generalized by Coifman and Meyer [2] as:

$$\begin{aligned}&0 \le \omega_j(t) \le 1 \quad \forall t \in \Re \\ &\omega_j(t) = 0 \quad t \le a_j - \alpha_j \text{ or } t \ge a_{j+1} + \alpha_{j+1} \\ &\omega_j(t) = 1 \quad a_j + \alpha_j \le t \le a_{j+1} - \alpha_{j+1} \qquad (6) \\ &\omega_{j-1}(a_j + \tau) = \omega_j(a_j - \tau) \text{ if } |\tau| \le \alpha_j \\ &\omega_j^2(a_j + \tau) + \omega_j^2(a_j - \tau) = 1 \text{ if } |\tau| \le \alpha_j\end{aligned}$$

And the expressions of the Malvar wavelets can be written in the two following forms:

$$u_{j,k}(t) = \sqrt{\tfrac{2}{l_j}} \cdot \omega(j) \cdot \cos\left[\frac{\pi}{l_j}\left(k + \tfrac{1}{2}\right)(t - a_j)\right] \qquad (7)$$

$k = 0,1,2,\ldots$ and $j \in Z$

or,

$$\begin{aligned}&u_{j,k}(t) = \sqrt{\tfrac{2}{l_j}} \cdot \omega_j(t) \cdot \cos\frac{k\pi}{l_j}(t - a_j) \quad \text{if } j \text{ even and } k = 1,2,\ldots \\ &u_{j,k}(t) = \sqrt{\tfrac{1}{l_j}} \cdot \omega_j(t) \quad \text{if } j \text{ even and k } = 0 \qquad (8) \\ &u_{j,k}(t) = \sqrt{\tfrac{2}{l_j}} \cdot \omega_j(t) \cdot \sin\frac{k\pi}{l_j}(t - a_j) \quad \text{if } j \text{ odd and } k = 1,2,\ldots\end{aligned}$$

The functions $u_{j,k}(t)$ form an orthonormal basis of $L^2(R)$. The general elementary wavelet is composed of an onset part, a stationary part and a release part. It is possible to modulate this three degrees of freedom via the parameters lj, α_j and α_{j+1} to create an optimised basis.

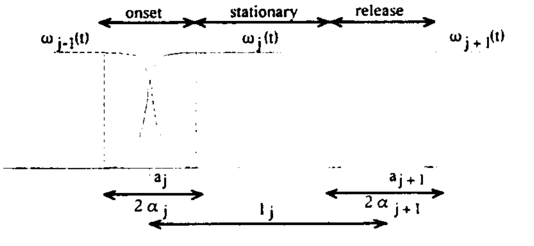

fig. 1 : definition of the overlapping window

Moreover, for each value of k, the expressions (7) or (8) represent the impulse response of a FIR, so that the set of wavelets forms a filter bank that covers the entire spectrum. As the representation of the signal is obtained by projecting it on the orthonormal basis $(c_{jk} = \langle u_{jk}, s \rangle)$, the coefficients c_{jk} provide a local spectrum of the signal and, due to the particular symmetry properties of this transformation, represent it in a complete and non-redundant way.

3. SPEECH PROCESSING

3.1 The Malvar cepstrum

The properties of the Malvar analysis and more particularly, its easy adaptation to variable length time-partition of any signal let foresee some opportunity for the automatic segmentation of the speech signal. Therefore, it is necessary to adapt the speech analysis to this segmentation in order to achieve coherent measurements whatever the length of a segment is. We will take advantage of the particular orthogonality properties of the Malvar wavelets to adapt the speech analysis. Indeed, these properties make comparable the measures achieved in the subspaces defined by the orthonormal basis of wavelets [2]. Actually, we are especially interested in fitting the feature extraction.

Since the coefficients of the Malvar Wavelet decomposition provides a local spectrum of the signal, we can apply the definition of the cepstrum to these coefficients and compute a Malvar cepstrum to represent any frame.

The Malvar cepstrum is thus defined as :

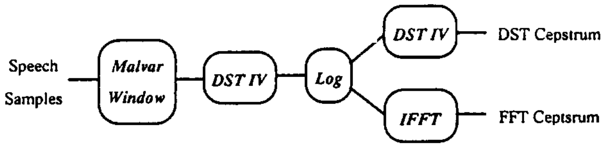

Fig. 2 : computation of the Malvar cepstrum

Results obtained with that kind of features will be presented later.

3.2 The segmentation algorithm

The speech signal is a non-stationary process, but it is assumed to be "quasi-stationary". That is, we can consider that over a short period of time the statistical properties of speech features are rather constant. Consequently, the feature extraction is usually achieved on fixed length segments short enough to allow the quasi-stationary assumption of the signal (about 10 ms). Actually, the notion of stationarity evolves along the signal, the quasi-stationary periods are longer in voiced speech than in unvoiced speech. So, a better approach would consist in developing a variable length segmentation of the signal into quasi-stationary parts.

Some expected advantages of such a segmentation in the framework of speech recognition task are :

- It seems that non-stationary parts are more important than stationary parts in speech recognition. A segmentation into variable length quasi-stationary units emphasise non-stationary parts since the density of acoustical vectors is higher in non-stationary parts than in stationary parts [6, 7].

- Using a HMM recognizer, we can expect a better modelling of state duration since the negative exponential form of the probability to stay *d* times in the same state is inappropriate for speech modelling especially for large values of *d*.
- Using ANN recognizers, the segmentation will provide an automatic and clever time-alignment of the speech signal.

The Malvar Wavelet decomposition can be used to obtain this segmentation. The principle consists in computing an optimal basis that maximises the information contained in the Malvar coefficients. This is achieved by minimising a value called the entropy which is defined by :

$c_k = \langle s, u_k \rangle$ s is the signal and u_k are the wavelets.

$$H(s) = -\sum_k |c_k|^2 \cdot \log |c_k|^2 \qquad (9)$$

The algorithm that leads to the optimal segmentation is based on the split & merge algorithm that consists in modifying a pre-existing segmentation. The operation of suppressing some of the existing points a_j, i.e. concatenating the segments $[a_{j-1}, a_j]$ and $[a_j, a_{j+1}]$ to the segment $[a_{j-1}, a_{j+1}]$, is called *merging*. Inversely the operation of inserting a point a_j in an existing segmentation is called *splitting*.

Now, let us see the effect of merging two segments on the orthonormal basis. We know that the basis of segments $[a_{j-1}, a_j]$ and $[a_j, a_{j+1}]$ span the subspaces W_{j-1} and W_j. Since the orthonormal basis are orthogonal to each other, merging the two segments is equivalent to replacing the subspaces W_{j-1} and W_j by their direct orthogonal sum. This is also equivalent to replace the windows $w_{j-1}(t)$ and $w_j(t)$ by a new window $\tilde{w}_j(t)$ defined as :

$$\tilde{w}_j(t) = \sqrt{w_{j-1}^2(t) + w_j^2(t)}$$

The actual segmentation algorithm is described by the two following steps :

- We create a set of fine-to-coarse dyadic time-partitions and, for each partition, we compute the Malvar coefficients taking care to adapt the window as suggested by the split and merge algorithm.

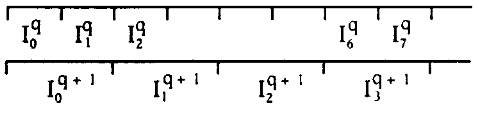

fig. 3 : set of pre-defined dyadic segmentations

- The concatenation of two segments obeys to the following entropy minimisation criterion :

$$\begin{cases} I_{2j}^q, I_{2j+1}^q & \text{if } H(\alpha_{2j}^q) + H(\alpha_{2j+1}^q) < H(\alpha_j^{q+1}) \\ I_j^{q+1} = I_{2j}^q \cup I_{2j+1}^q & \text{otherwise} \end{cases} \qquad (10)$$

q : level of segmentation

j : index of the segment

So doing, we obtain an adaptea segmentation which, due to the choice of an entropy criterion, leads to a time-partition of the signal into quasi-stationary units. Therefore, in case of speech signal, the segmentation is related to the phonemic evolution of the signal, that is, a sequel of short segments for unvoiced and transient speech and some large segments for voiced speech.

The entropy can be interpreted as being a measure of the number of significant coefficients of the decomposition. This can be understood by analogy to entropy defined in Information Theory. Information Theory defines the entropy as [8]:

$$h = \sum_i p(x_i) \log p(x_i) \qquad (11)$$

where $p(x_i)$ is the probability density function of x

Entropy in this case represents the amount of information present in the signal x. If $p(x)$ is a uniform distribution, h is maximum. At the opposite, if $p(x_i)=0$ for all x_i except x_1 and $p(x_1)=1$, then $h=0$ and x does not contain any information.

In equation (9) coefficients $|c_k|^2$ play the same role as the probabilities in (11). So we can interpret equation (10) in the same way as equation (11) :

Equation (11)	Equation (9)
h is max. if p(x) is a uniform distribution	H is max. if $\lvert c_k \rvert^2$ are constant (flat spectrum)
h decreases if p(x) strays from uniform distribution	H decreases when formants appear in the signal

Table 1 : interpretation of the entropy

So, during stationary periods, since formants are added while noise components are averaged, the entropy on a long segment is lower than the sum of the entropies of the two small segments. On the other side, during noise periods or transient parts, the averaging effect will flatten the spectrum so that entropy of long segments is higher than the sum of the entropies of the short segments.

Here are some examples of segmentation obtained with 4 levels of segmentations from 64 to 512 samples.

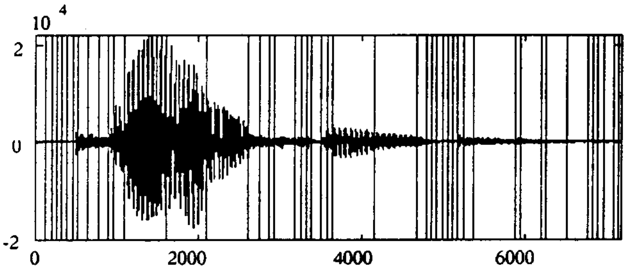

fig. 4 : "tausend"

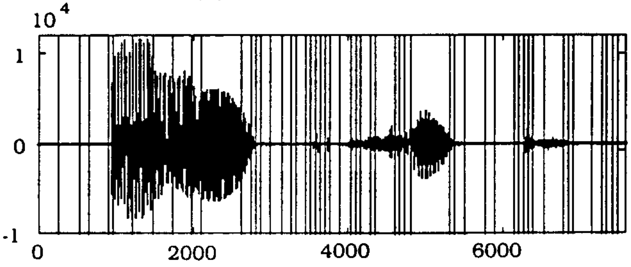

fig. 5 : "neunzig"

3.3 Feature extraction

The feature extraction must be adapted to the variable length segmentation in order to emit only one acoustical vector by segment. Much care has to be taken for this task not to make illegal things. Indeed, for instance, due to differences in the form of the windows, the measurements issued from a STFT (as LPC cepstrum and FFT cepstrum) can not be compared between segments of different lengths. A quite sure way to compute the acoustical vectors consists in, first, achieving the feature extraction on the finest level of segmentation and then calculating the mean vector on the larger segments. Nevertheless, this trivial approach does not take advantage from the fact that the selected sequel of segments is optimal in terms of the information quantity contained in the coefficients of the Malvar decomposition. Another way to carry out the feature extraction consists in computing the vectors in subspaces in which the measurements are comparable and coherent whatever the length of a segment is. This is possible using the Malvar cepstrum we defined in paragraph 3.1.

4. RESULTS

We have tested the pre-treatment proposed here in an isolated word speech recognition task, on telephone line (53 English words, 190 speakers for the training set and 36 speakers for the test set). The first results deal with the use of the Malvar cepstrum on fixed length segmentation in order to verify if there is some sense to use them in recognition task. The following results concern comparison of speech recognition on variable length segmentation first using discrete phoneme model HMMs, then using word model HMMs.

Model 1 : 12 cepstra - phoneme models HMM
Model 2 : 12 cepstra - cms - phoneme models HMM
Model 3 : 12 cepstra, 12 Δ-cepstra, Δ-energy and
$\Delta\Delta$ -energy, cms, word models HMM

	LPC Cepstrum	FFT Malvar Cepstrum	DCT Malvar Cepstrum
model 1	67.64 %	59.82 %	58.29 %
model 2	76.44 %	79.58 %	77.07 %
model 3	94.31 %	94.03 %	-

Table 2 - Recognition rates with Malvar cepstra

We can conclude from these experiments that the Malvar analysis is well suited for speech processing. Note that the use of a classical convolution noise suppressing technique (CMS [9]) results in a very high improvement of the recognition rates.

	without segmentation	with segmentation
Cepstra (1 codebook + CMS)	81.61 %	78.71 %
Malvar FFT (1 codebook + CMS)	84.02 %	76.74 %

Table 3 - Recognition rates on variable length segmentation

In all the cases, the recognition rates were better without segmentation. It seems that we used it badly with HMM. Maybe because the HMM are well suited to model stationary parts while we emphasise non-stationary parts. This weakness can be avoided by using a segment based recognition algorithm instead of HMMs.

5. CONCLUSIONS

We have proposed in this paper a new pre-processing algorithm for speech signals based on Malvar wavelets producing a new type of acoustical vectors at a variable frame rate according to the stationarity of the signal. Improvement was observed with the Malvar cepstrum when noise compensation technique is performed. This results seem to show that the Malvar cepstrum is particularly suited for noise compensation. Moreover, our algorithm present a big interest for recognizers that need some time-alignment and that focus on transition parts like neural networks.

6. ACKNOWLEDGEMENTS

The authors would like to thank H. Bourlard from ICSI Berkeley, J. Vantieghem and S. Accaino from the industry Lernout & Hauspie Speech Products and J. Hennebert from Federal Institute of Technology of Lausanne for their contribution to this study. This work is supported by the Commissions of the European Community through the ESPRIT Project 6488 HIMARNNET.

7. REFERENCES

[1] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis", IEEE Transactions on Information theory, vol. 36, Sept. 1990, pp 961-1005.

[2] Y. Meyer, "Les Ondelettes, algorithmes et applications", Editions Armand Colin, 1992.

[3] E. Wesfreid and M. V. Wickerhauser, "Adapted Local Trigonometric Transforms and Speech Processing", IEEE Trans. on Signal Processing, December 1993, pp. 3596-3600.

[4] H. S. Malvar, "Lapped Transforms for Efficient Transform/Subband coding", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 38, 1990, pp. 969-978

[5] H. S. Malvar, "Signal Processing with Lapped Transforms", Artech House Edition

[6] N. Morgan, H. Bourlard, S. Greenberg, H. Hermansky, "Stochastic Perceptual Auditory-Event-Based Models for Speech Recognition", Intl. Conf. on Spoken Language Processing (ICSLP), Yokohama, Japan, September 94.

[7] O. Ghitza and M. M. Sondhi, "Hidden Markov Models with Templates as non-stationary states : an application to speech recognition", Computer Speech and Language, Vol. 7, Number 2, pp. 101-119, April 1993.

[8] T. M. Cover & J. A. Thomas, "Elements of Information Theory", Wiley Interscience Edition, 1992.

[9] C. Mokbel, J. Monne, D. Jouvet, "On-line Adaptation of Speech Recognizer to variations in Telephone Line Conditions", EUROSPEECH 93, Vol.II, pp 1247-1250