Application of the wavelet transform for speech processing.

Stéphane Maes
(ISU'91-ISU'93-BEL)

Rutgers University,
Center for Computer Aids for Industrial Productivity (CAIP),
P.O. Box 1390, Piscataway, NJ 08855-1390, USA.

ABSTRACT

Speaker identification and word spotting will shortly play a key role in space applications. This paper presents an approach based on the wavelet transform. In the context of the "Modulation Model" [1], enables to extract the speech features which are used as input for the classification process.

1. INTRODUCTION

Speech processing has always played an important role in telecommunications. The primary goal is to transmit, or store, efficiently speech signals. It quickly appeared that synthesizer-coders are among the most efficient [2]. Therefore, huge efforts have been made to develop speech production models and to come up with features sets and models which completely characterize a signal of speech. This paper describes a technique based on the wavelet transform which allows on line extraction of such a set of parameters. Those parameters proved well adapted for speaker identification and word spotting, in clean as well as in a noisy environment.

Those excellent results open new perspectives for security systems and for voice controlled devices, which have huge potential for space activities. Robust voice control systems alleviate a good many ordinary task. They increase the reaction speed and efficiency of operators involved in tricky maneuvers. Voice commands increase also the number of degrees of freedom which can be simultaneously modified, by freeing the operator's hands even by allowing the operator to move freely in the control room. This requires the use of microphone array with directionality controlled by a speaker identification system, in order to follow the right operator in a noisy and even crowded room. Coupled with virtual reality (VR) techniques, voice control systems provide powerful and more user friendly interfaces for remote control of robotics tasks.

2. MODULATION MODEL

Speech results from the excitation of the vocal tracks at a fundamental frequency called the pitch [1,2,3]. The resulting signal produced by the vibration of the vocal tracks and composed by a fundamental component and some of its harmonics is modified (filtered) during its transmission from the vocal tracks till the lips. The transfer function which characterize this modification takes into account the length of the path, the shape of the vocal channel (inner mouth, tongue, glottal constriction,...) and the presence of resonators (e.g. nasal pits,...). The pitch and the transfer function are time dependent.

Based on those speech production considerations, the "modulation model" [1] considers that speech signals can be written as linear combination of principal components, each being characterized by an instantaneous amplitude ($A_i(t)$) and an instantaneous phase ($\phi_i(t)$):

$$s(t) = \sum_{i=1}^{N} A_i(t)\cos(\phi_i(t)) + \eta(t) \qquad (1)$$

where $\eta(t)$ denotes the model error due to the finite summation plus additional noise.

In order to perform efficient speaker identification or word spotting, it is mandatory to extract adequate features of the speech signal prior to any classification process based on them (see Figure 1). Until now, the LPC derived cepstrum (linear prediction coefficients) approach gives the best results. This can be understood in the context of the "modulation model". In fact, LPC considers the speech as a piece wise

stationary signal approximated by a linear combination of principal components, with instantaneous amplitude:

$$A_i(t) = a_0 e^{-\sigma_i t} \tag{2}$$

and instantaneous phase:

$$\phi_i(t) = \omega_i t + \theta_i \tag{3}$$

In the absence of noise, it can be proven [4] that the cepstra $c_n(t)$ behave as new signals with $A_i(t)$ and $\theta_i(t)$ forced respectively to $\frac{1}{n} e^{-\sigma_i(t)n}$ and 0:

$$c_n(t) = \frac{1}{n} \sum_{i=1}^{N} e^{-\sigma_i(t)n} \cos(\omega_i(t)n) \tag{4}$$

The cepstrum approach is sensitive to noise and contains only a partial amplitude information (e.g. the bandwidth σ). Improvements are expected if more general techniques are used to determine the $A_i(t)$ and $\phi_i(t)$ without the kind of restrictions imposed by the LPC analysis.

Most of the techniques (e.g. em Hilbert transform, Wigner-Ville, Choi-Williams) which have been proposed so far to separate the different components and extract their instantaneous amplitudes and frequencies are time consuming, difficult to implement and nonlinear which implies also a high sensitivity to noise [5].

3. THE WAVELET TRANSFORM

Window Fourier transform (or Gabor transform) and wavelet transform are efficient alternatives.

The window Fourier transform consists into a Fourier transform of the signal pre multiplied by a well chosen window:

$$g_{(b,\omega)}(x) = e^{j\omega(x-b)} g(x-b) \tag{5}$$

$$G_f(b,\omega) = < g_{(b,\omega)} | f > \tag{6}$$

It satisfies the identity reconstruction:

$$f(x) = \frac{1}{2\pi < g|h >} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} G_f(b,\omega) h_{(b,\omega)}(x) d\omega db \tag{7}$$

The time-frequency support of such an analysis remains the same at each frequency level. This leads to problems whenever the signal exists on a wide range of scales, which is usually the case for speech.

The wavelet transform is defined by the following equations:

$$g_{(b,a)}(x) = \frac{1}{a} g\left(\frac{x-b}{a}\right) \tag{8}$$

$$W_f(b,a) = < g_{(b,a)} | f > \tag{9}$$

It satisfies the identity reconstruction:

$$f(x)=\frac{1}{C_{(g,h)}}\int_{0}^{+\infty}\frac{da}{a}\int_{-\infty}^{+\infty}dbW_f(b,a)h_{(b,a)}(x) \qquad (10)$$

provided that the admissibility condition is satisfied:

$$0<C_{g,h}=\int_{0}^{+\infty}\frac{\hat{g}(\omega)\hat{h}(\omega)}{\omega}d\omega<+\infty \qquad (11)$$

There are two different ways to visualize the wavelet transform, as illustrated in Figure 2. Firstly, it is the correlation value between the signal and dilated and translated versions of a band pass function: the wavelet. Secondly, it can also be considered as the output of a filter bank where each filter has a dilated version of the wavelet as impulse response. Due to the dilation process, the time-frequency support depends on the frequency location: low frequency events are studied on large time intervals (with narrow frequency support), while high frequency contents are studied on narrow time intervals (with large frequency support). See Figure 3.

Both the Gabor and the wavelet transforms provide perfect reconstruction formula, which, in the case of the wavelet transform, defines the perfect reconstruction synthesis filter bank.

Also as those transformations of one dimensional functions are defined in two dimensional spaces, they are over complete. Therefore, they can be sampled [5]. Critically sampled wavelet transforms constitute particular cases of subband coding decompositions, a well known signal processing technique. In subband coding, a signal is successively split up into two subbands. One is roughly the low frequency content of the input signal (i.e. it is obtained with a low pass filter) and the other is the detail signal needed to reconstruct the original signal. The different subbands are downsampled by a factor two. The decomposition is iterated on the low frequency subband. The reconstruction filter bank is designed to recombine the two subbands and to cancel the aliasing errors introduced by the downsampling operation. Any other type of subband coding decomposition can be related to an hybrid wavelet transform where additional operators are applied on the wavelets [6,9,10].

To allow on line quasi-continuous wavelet transforms, a fast wavelet transform algorithm has been developed. It is shown in Figure 4. It is based on an extension of the classical subband coding approach to a quasi-continuous situation and is particularly well adapted to parallel and hardware implementations. It consists in a projection upon an intermediate set of functions called scaling functions, followed by a decomposition of the scaling function into the wavelet functions, instead of a direct projection upon the set of wavelets [7,10].

4. PARAMETER EXTRACTION

4.1. The ridge skeleton algorithm

Once the wavelet transform is computed, different algorithms can be used to extract the speech features $A_i(t)$ and $\phi_i(t)$. The first technique is known as the ridge-skeleton approach. It is based on the steepest descent method approximations which single out the dominant contributions to the wavelet transform, to which it associates curves into the time/scale plane(i.e. the ridges). An example is presented in figure 5. When all the hypotheses, needed to perform these approximations, are satisfied, the ridges provide the phase modulation laws of the components while the restriction of the wavelet transform to the ridges (i.e. the skeleton) enable us to evaluate the amplitude modulation laws. Unfortunately, these hypotheses are not always satisfied and, in that case, only sophisticated perturbation techniques sometimes give correct results [8].

4.2. The fusion and squeezing processes

We have proposed the fusion approach based on the filter bank concept. The signal is split up into different frequency subbands. The temporal evolution of some properties of these subbands can easily be monitored. The subbands which have the same temporal behavior are recombined with the synthesis filter bank. Its outputs are the principal components of the original signal. this is sketched in Figure 6. As criterion for recombination, and in order to provide a practical time-frequency representation, we have introduced a squeezing process. It consists into building a measure function ($\mu(b,\omega)$), defined into the time/frequency plane. It is, at a given time location, a function of the frequency, which is obtained by adding together the wavelet transform (or only its modulus, or a linear combination of the wavelet transform and its modulus) of all the different subbands which have a contribution at this frequency. A typical squeezed plane is presented in Figure 7.

$$\omega(b,a) \therefore = \partial_b \phi_{w_f}(b,a) \tag{12}$$

$\phi_{w_f}(b,a)$: phase of the wavelet transform.

$$\Omega = [\omega - \delta_\omega, \omega + \delta_\omega] \tag{13}$$

c(a) is a weighting factor.

$$\mu(b,\omega) = \left| \int_{\omega(b,a)\in\Omega} (|W_f(b,a)| + W_f(b,a)) c(a) da \right| \tag{14}$$

This method uses the whole information present into the time/scale plane, without any approximation. Into the squeezed plane, if we take only the modulus of the wavelet transform, otherwise, into the synchrosqueezed plane, the instantaneous frequencies of the principal components can be tracked by dynamic programming techniques [7].

The squeezed and synchrosqueezed plane presents the phase modulation of the different components. Once those components are identified, they can easily be reconstructed with the following strategy which consists into a simple addition. This is what we call the the fusion algorithm:

$$\int_0^{+\infty} \frac{da}{a} W_f(b,a) = f(b) A_g \tag{15}$$

$$A_g = \int_0^{+\infty} \frac{du}{u} \hat{g}(u) \tag{16}$$

Those components have a known angular modulation. Therefore it is easy to extract the amplitude modulation of each component.

The wavelet approach (as well as the Gabor approach) is almost completely linear, and the only non linearity results from the measure computation, which is an averaging process, followed by dynamic programming. This guarantees good robustness to noise. In fact for a Signal to Noise Ratio: SNR=10dB, it is still possible to obtain a very good extraction of the pitch and of the first three formants. This has to be compared with the threshold of 20dB of the other classical formant extractors.

Figure 8 and figure 9 present the synchrosqueezed plane obtained for two speech signal in a noisy environment. The different formants are clearly visible.

Some inaccuracies might also occur for the higher frequencies due to a wide frequency support which prevents different components from being resolved by the wavelets. However, this can be easily solved with wavelet packets, based on a similar approach, where the wavelets are scale adapted.

5.CLASSIFICATION

This feature extraction is a front-end for the classification process. Different techniques [2] exist for feature classification of speech. The most efficient ones are (presented in the case of speaker identification:

- Dynamic Time Warping. The time dependent set of feature stretched and deformed by a mapping function, in order to match each speaker in the data base. The identified speaker is the one who has the mapping function with the lowest cost (Figure 10).
- Hidden Markov Models. Each speaker is characterized by a set of "states" and probability of transition between those states (Figure 11). The probability to obtain the observed feature set is computed for each model and the one with the highest probability is considered as the identified speaker.
- Vector Quantization: A dictionary of dominant feature pattern is produced for each speaker. The observed feature set is coded with each dictionary. The most efficient code identifies the speaker.
- Autoregressive Vector Models. Each speaker is characterized by a autoregressive model. The probability to obtain the observed feature set is computed for each model and the one with the highest probability is considered as the identified speaker.
- Neural Networks and Neural Tree Networks. Neural networks perform direct classification by splitting up with hyperplanes, an hyperspace into different volumes. Each volume (sometimes more than one) is associated to a different speaker. Neural Tree Networks allow to increase the database without having to retrain the whole system. The classification is hierarchical. It starts with a very simple neural networks and each time that two speakers have to share the same volume in the hyperspace, an new simple neural network is used to discriminate between them.

Excellent speaker identification results have been obtained [1,4]. The error rate is lower than 2% with databases of more than one hundred speakers. The system is totally text and language independent. For example, it is possible to train the system in one language (e.g. Thai) and to test it in another (e.g. French). Channel distorsion are under investigation in order to increase the robustness to noise.

6. ACKNOWLEDGMENTS

Stephanie Maes is research assistant of the Belgian National Fund for Scientific Research (FNRS) at the Université Catholique de Louvain, Louvain-la-Neuve, Belgium: Laboratoire de Télécommunications et Télédétection (TELE) and Unité de Physique Théorique et Mathématique (FYMA). The author wishes to thank professor Ingrid Daubechies (ATT Bell Labs) for her useful help and support.

7. REFERENCES

1. K.T. Assaleh, R.J. Mammone, M.G. Rahim, J.L. Flanagan, "Speech recognition using the modulation model", Proc. ICASSP, 1993.
2. L. Rabiner, Biing-Hwang Juang, Fundamental of speech recognition, Prentice-Hall, 1993.
3. S. Furui, Digital Speech Processing, Synthesis and Recognition, Marcel Decker, 1987.
4. K.T. Assaleh, K.R. Farrell, R.J. Mammone, "Speaker recognition using modulation model parameters and neural network classifiers", submitted to IEEE Trans. on Speech and Audio, 1993.
5. I. Daubechies, Ten Lectures on Wavelets, SIAM, 1992.
6. S. Maes, " Adaptive Optimal Decompositions with Hybrid DCT-Subband Coding", Proc. IEEE -SP, Time-Frequency and Time-Analysis, (The IEEE Signal Processing Society, 1992), pp. 479.
7. S. Maes, " An Application of Wavelets to the Characterization of Formant Parameters in Speech.", ATT Bell Labs seminary, Holmdel, May 7, 1993.
8. N. delprat, E. Escudoc, P. Guillemain, R. Kronland-Martinet, Ph. Tchamitchan, B. Torresani, "Asymptotic Wavelet and Gabor Ananlysis: extraction of instantaneous frequencies", IEEE Trans. Inf. Theory, vol 32, n°2, part 2 , pp. 644-664, March 1992.
9. S. Maes, " Hints on the Relationships between Subband Coding, Wavelet and Vaguelette Decompositions and Cosine Sine Alterned Transformations", Proc. EUSIPCO, Elsevier, 1992.
10. S. Maes, "Hybrid Subband Coding Decompositions", ATT Bell Labs seminary, Murray Hill, May 18, 1993.

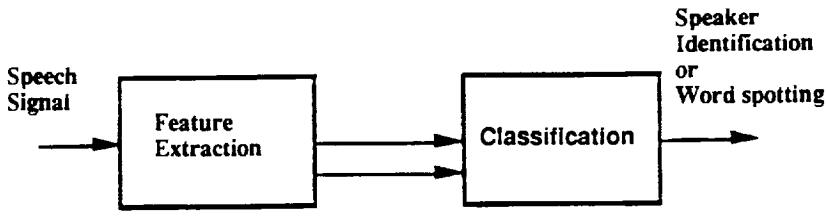


Figure 1: For speaker identification or word spotting, features have to be extracted prior to any classification process.

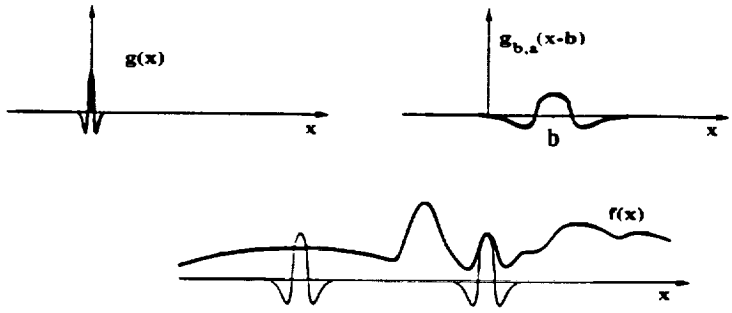


Figure 2: The wavelet transform, can be considered as the correlation value between the signal and a set of dilated and translated version of the "original" wavelet. Because of the admissibility condition, this transform peaks when signal and wavelet looks alike and it is negligible everywhere else.

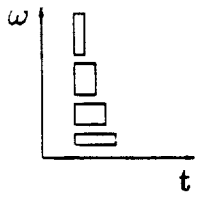


Figure 3: The time frequency support of a wavelet analysis changes with the frequency location.

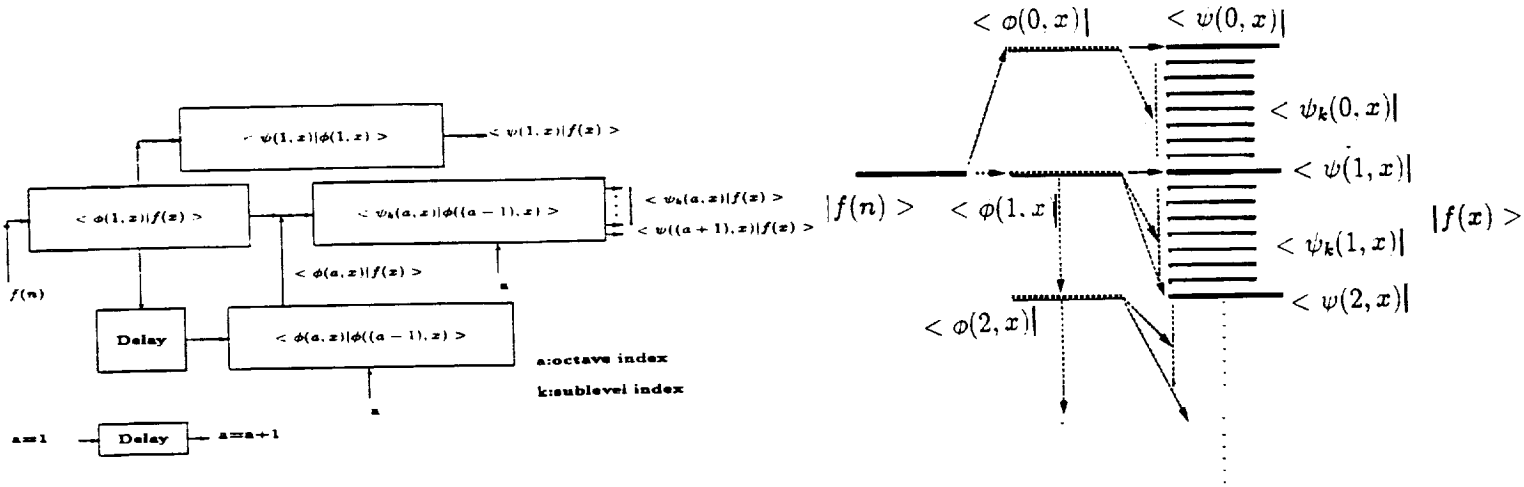
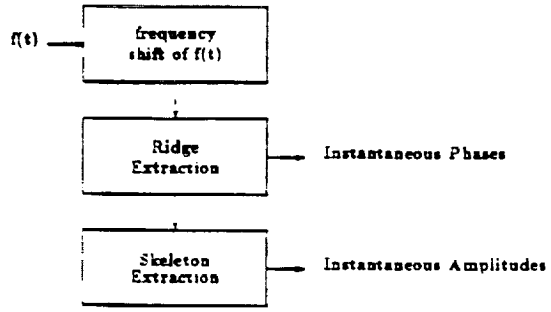


Figure 4: It describes the "indirect" computation of the wavelet transform. It is extremely fast because it can be implemented with a few filter banks.



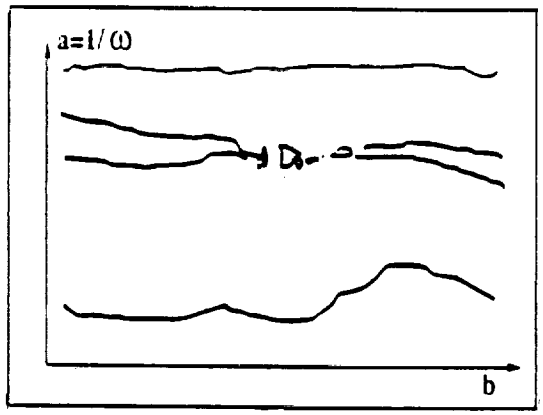


Figure 5: It describes the ridge-skeleton process and presents a typical ridge in the time/scale plane.

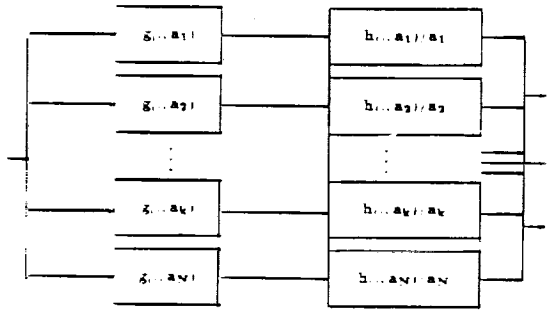


Figure 6: The fusion process. Every subband which presents the same temporal behavior are recombined. The output are the different components of the signal.: the formants.

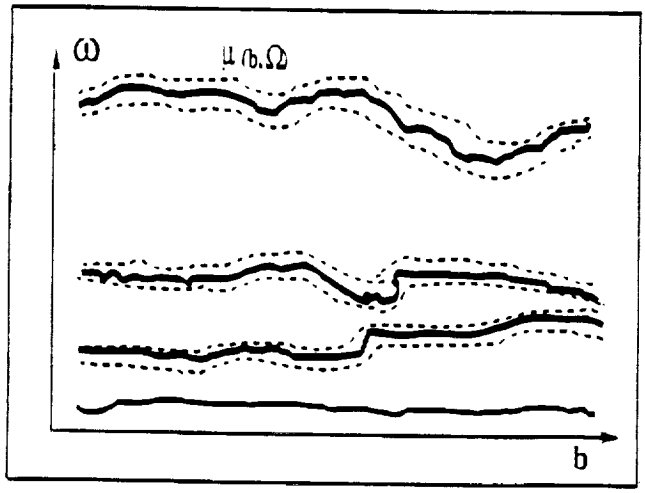


Figure 7: A typical synchrosqueezed plane. The peak values define the central frequency while the bandwidth is easily determined by relative threshold.

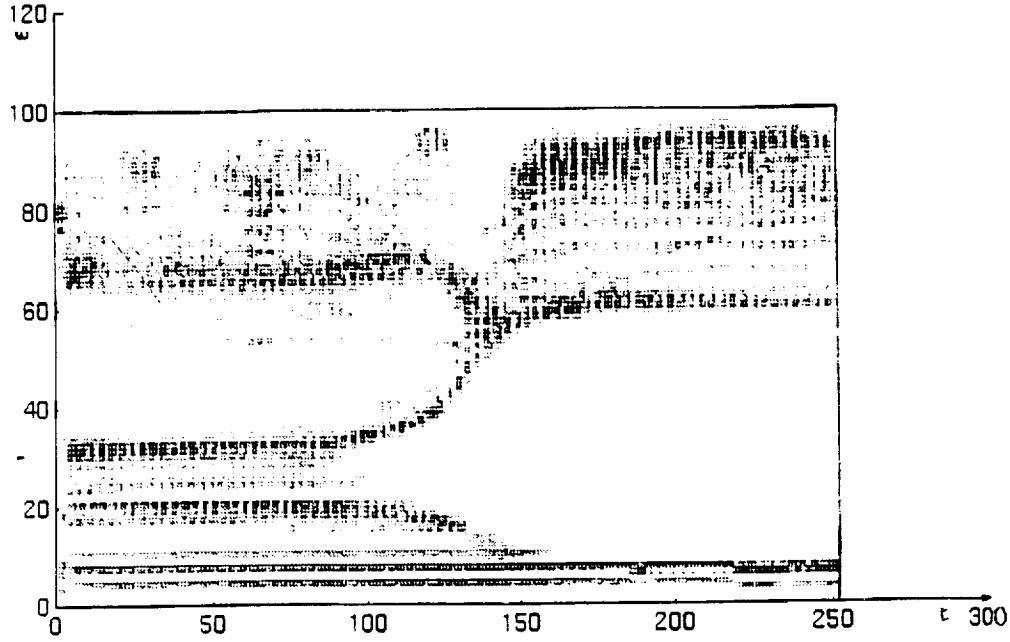


Figure 8: The synchrosqueezed plane obtained for the transition in "aaii" (French pronunciation) said by a male voice, with a SNR=15dB, pink noise. The frequency scale is linear from 0 to 4000Hz. The time interval is roughly 0.5 s.

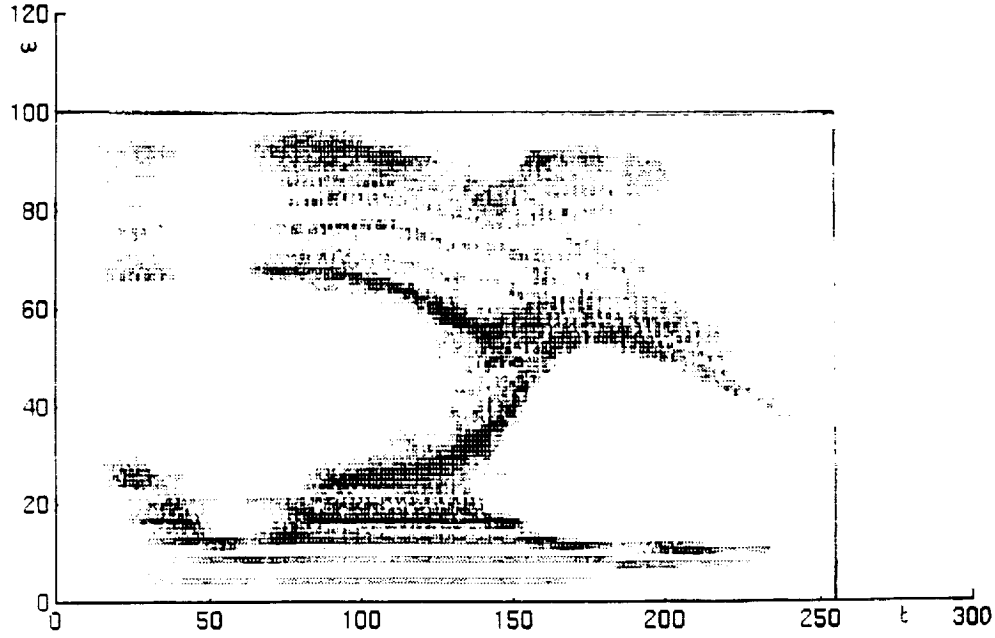


Figure 9: The synchrosqueezed plane obtained for "how are you?" said by a male voice, with a SNR=15dB, pink noise. The frequency scale is linear from 0 to 4000Hz. The time interval is roughly 0.5 s.

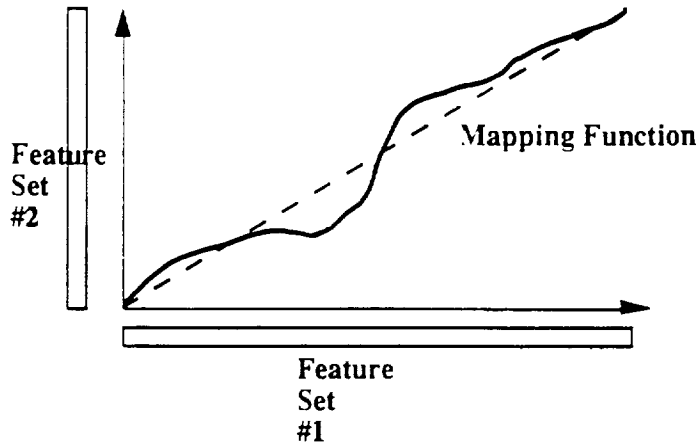


Figure 10: It illustrates the definition of the mapping function for the Dynamic Time Warping approach.

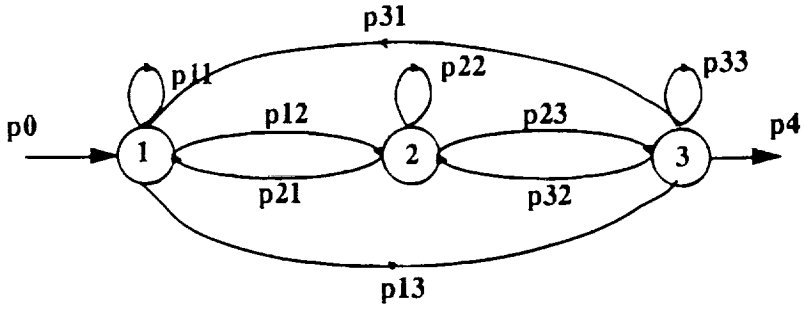


Figure 11: It illustrates the Hidden Markov Model approach. p_{ij} denotes the probability to have a transition from the state i to the state j, knowing that the system is in the state i.

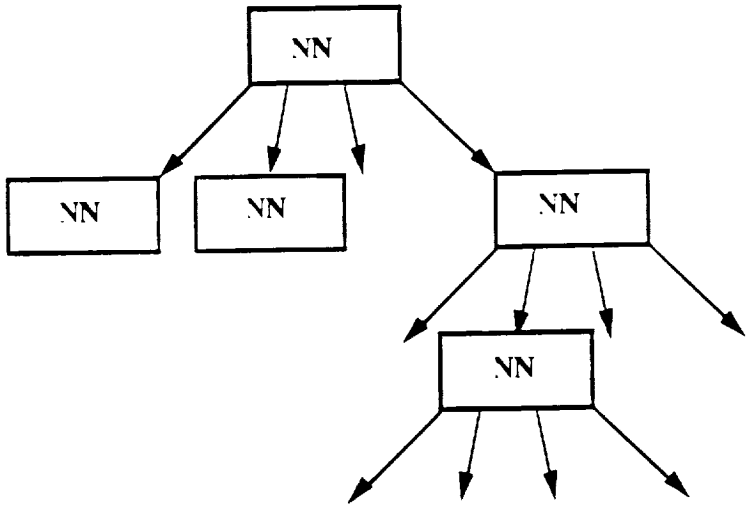


Figure 12: It illustrates the concept of Neural Tree Networks.