ALTERNATIVE SPEECH REPRESENTATIONS FOR KOHONEN CLASSIFIERS

A. P. Reilly and B. Boashash

CRISSP, Department of Electrical Engineering,
The University of Queensland

ABSTRACT - Several different techniques have been used to process speech signals in preparation for classification. The most commonly used ones rely on assumptions of signal stationarity that are not true for many important speech signal types. Recent work in the field of time-frequency signal analysis has produced a number of representations which do not make these restrictive assumptions. This paper reports on work being done to quantify the differences between these representations, as relates to the classification of speech signals using the neural network architecture popularized by Teuvo Kohonen in 1984.

INTRODUCTION

A typical speech recognition system has three distinct sections (see figure 1, below): a data acquisition section which translates sound waves into a form suitable for processing, a filtering, or spectrum extracting section, which makes repetitive, frequency-domain features of the signal apparent, and a classification section, which labels the speech signals in some meaningful way.

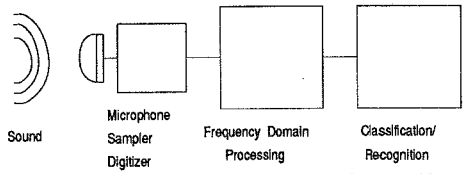

Figure 1. Block representation of mechanical speech recognition system

In mammals, the acquisition task is performed by the outer and middle ears, which convert air pressure variations into small mechanical vibrations. The spectrum estimation task is performed by the cochlea, which produces nerve impulses relating to the frequency content of the sound. The classification process is performed by the relevant sections of the brain.

In digital systems the data acquisition system usually involves a microphone, an amplifier, an anti-aliasing filter, and an analog-to-digital converter/sampler. The last two sections are usually performed by digital computers, using a number of different techniques. Commonly the spectrum estimation task is performed using some form of Fourier transform, linear predictive coding, or cepstrum estimation, while the classification task has been performed variously by matched-filter templates, dynamic time warping, expert systems, hidden Markov models, and neural network models.

The speech signal is characterized by periods of fairly stationary frequency spectra, with multiple energy peaks corresponding to the 'formants' or resonances of the voiced sounds. These are punctuated by regions of ramping frequency (the liquids and semi-vowels), periods of coloured noise (the fricatives), periods of silence (the glottal stops), and very short bursts of noise energy (the plosive releases). One of the main failings of automatic speech recognition systems to date has been in the area of recognizing the less stable speech sounds. That is, everything other than the vowels. This is a significant shortcoming, because it is believed that a lot of the speech information is carried in these transient signals.

It seems pointless to embark on a project to produce a speech recognition system based on an input representation known to be unable to represent some speech signals. As has been mentioned, this is the case with most of the techniques popular today. It is necessary, therefore, to investigate other speech representation options, which might be capable of more complete information transfer.

In the following section a review of the relevant characteristics of some of the standard speech representation techniques will be presented, along with the characteristics of some of the newer techniques.

TIME-FREQUENCY REPRESENTATIONS REVIEW

This review will look at those aspects of the selected representations that intuitively bear on their usefulness as a speech signal coding.

- A successful speech representation should not miss the rapid frequency changes found around plosives and in some liquids.

- The representation should be stable when frequency components merge, cross, start and stop. This is largely so that 'continuing' components are recognized as such. Parametric representations tend to jumble the order of features at discontinuities, making this difficult.

- Fine time resolution is probably important for classifying some of the plosive release sounds.

- It would be nice to be able to avoid the artifacts produced by most bilinear time-frequency representations, such as the Wigner-Ville distribution, when used to analyze multi-component signals. This may not be a real constraint, but the presence of time-frequency signals that do not correspond to time signals seems intuitively unacceptable.

- Although it should not be necessary to replicate the performance of the human ear to perform speech recognition, this *is* an example of a working speech recognition system, and it would perhaps be a good idea to approximate its characteristics.

Short Time Fourier Transform:

The short-time Fourier transform (STFT) is the periodogram spectrum estimate of the product of the time signal with a window centred at a number of (usually evenly spaced) instants in time (see Rabiner and Gold, 1975). It is very widely used, and forms the basis for the familiar 'sonogram' representation of speech.

In practice, it is calculated at a fixed number of evenly spaced frequencies, using one of the fast implementations of the discrete Fourier transform (a Fast Fourier Transform or FFT), with a window whose length is on the same order as the principle pitch period, in order to achieve sufficient frequency resolution to locate the voiced formants. An unfortunate side-effect of this restriction on analysis window length is that it also limits the time resolution to the same order of size. This means that transient signals, such as plosive releases may be missed or misinterpreted as a result of being smeared into the other signals occurring during the window period.

Auto-Regressive Model Spectrum Estimate:

The Auto-Regressive (AR) Model spectrum estimate (see e.g., Marple, 1987), is one of the class of "modern" spectrum analysis techniques whereby assumptions about the structure of the signal are used to allow sharper resolution of the peaks of the spectrum. In this case the assumption is that the signal can be modelled accurately by an auto-regressive model with a finite number of parameters. That is, a model whose output is a linear combination of the input signal and previous values of the output signal. Another way of describing it is as an all-pole filter. This is actually a surprisingly good approximation for most of the vowels. Its failings are that the nasal sounds are best modelled by a filter containing zeros (which an AR model can only approximate), and that most of the parameter estimation techniques assume signal stationarity over some moderately long analysis window, just like the short-time Fourier transform.

Reflection Coefficients:

An efficient way to calculate the Auto-Regression (AR) coefficients used for the previous method is

the Durbin-Levinson Recursion algorithm, (Marple, 1975) which fits successively larger models to each block of speech data, up to some pre-determined maximum model order. The last model coefficient calculated for each order is referred to as a 'reflection coefficient', and has some physical significance. They can be related to the sizes of resonant cavities, and are sometimes used in remedial speech programs to provide some indication of the positions of various parts of the vocal tract.

These parameters should vary continuously for continuously varying speech sounds (although the actual AR coefficients do not), and they convey the same information as the AR model power spectrum, more compactly. They might therefore be at least as useful a speech representation as the AR model spectrum, and possibly more efficient.

Wavelet Transform:

The wavelet transform is a time-scale rather than a time-frequency transform, but its output can usually be given a time-frequency interpretation (see Daubechies, 1990). Where each of the discrete Fourier transform coefficients can be thought of as the projection of the signal onto one of a set of frequency-shifted sinusoids, the wavelet transform is the projection of the signal onto a set of time-scaled versions of an analysing function called a wavelet. If the wavelet has some oscillatory nature, then the longer wavelets will 'select' the lower frequencies, and the shorter ones the higher frequencies. A typical wavelet might look like the one shown in figure 2:

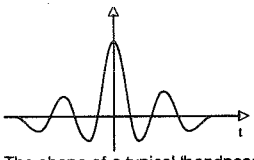

Figure 2: The shape of a typical 'bandpass' wavelet.

This should have good localization in both time and frequency, and may be complex. In the interests of convenient inversion, certain constraints are placed on the functions that can be used but a commonly used one is based on a Gaussian-weighted complex sinusoid.

Another way to think about the difference between the short-time Fourier transform and the wavelet transform is in terms of filter banks. The short-time Fourier transform can be considered to be the output of a bank of band-pass filters centred on evenly spaced frequencies, each with the same impulse response length, and the wavelet transform can behave like a bank of filters with constant Q, centered on a geometric progression of frequencies. This means that time resolution is traded against frequency resolution: high frequencies can be resolved with high time resolution, but low frequency resolution, while the opposite is true of low frequencies. This is conceptually nice, and actually quite similar to the response of the cochlea.

The wavelet transform is not limited to behaving in this way, as the 'analyzing wavelet' might be chosen to exhibit other properties, but wavelets can be designed to exhibit the nice properties mentioned. Longer wavelets can give better frequency resolution at the expense of time resolution.

For analyzing wavelets meeting the restrictions which classify them as 'orthonormal bases' (see Daubechies 1990,) the wavelet transform can be implemented very efficiently, based on trellis filters, similar to those used for subband coding of speech and video images. This gives the implementation a performance on the same order as that of the discrete Fourier transform.

The Bilinear Time-Frequency Distributions:

A relatively new type of signal analysis technique is based on the members of Cohen's class of bilinear distributions (Cohen, 1989). It turns out that the spectrogram can be described as a member of this class, but for this discussion it is probably best left where it is. The members of this class are

truly time-frequency representations, with no restrictions on the stationarity of the signal. It is possible to construct members of the class with the following desirable properties:

- The integral over frequency at any time produces the instantaneous energy of the signal.
- The integral over time at any frequency produces the stationary power spectrum.

Another characteristic of these representations is that they will resolve the frequency of signal components whose frequencies are rapidly changing.

A disadvantage of them, and a direct consequence of their product-kernel formulation, is that spurious peaks can appear between the different components of a signal. This has largely been solved now. (Zhao, et. al., 1990, and Choi & Williams, 1989). Three members of this class will now be described.

Windowed Wigner-Ville Distribution:

The Wigner-Ville Distribution (WVD) (see Boashash, 1990) is the simplest of the distributions in Cohen's class, as it is the 'default' case. All of the other distributions are filtered versions of the WVD. This gives it some advantages over the others and some disadvantages.

The WVD has a number of nice properties for signals which are essentially linear-FM chirp signals, relating to resolution and instantaneous frequency determination. Unfortunately, for signals (such as speech) where multiple frequency components are present at once, the WVD produces a 'cross term' at the mean frequency of each pair of components. This can make the representation difficult to interpret, but it is not certain that they would hinder a mechanical pattern recognition system.

Choi-Williams Distribution:

The Choi-Williams Distribution (see Choi & Williams, 1989,) is one case of Cohen's class of time-frequency distributions. It has a parameter, usually referred to as 'sigma', which affects the amount of filtering performed. The filtering suppresses the cross-terms, while reducing the resolution of the self-terms. In one limit, the distribution reduces to the Wigner-Ville distribution. In the other limit the frequency resolution is the same as that of the STFT, but the time-marginal condition (see above) is still maintained, so transients can be accurately positioned in time.

Zhao-Atlas-Marks, or Cone-Kernel Distribution:

This distribution (see Zhao et. al., 1990) is another member of Cohen's class of distributions. As well as suppressing the cross-terms caused by the bilinear nature of the kernel, it positions them *on top of* the self-terms. This means that there aren't any spurious peaks between components of the distribution. Instead, the peaks representing the actual signal components oscillate in amplitude. It is mooted by Zhao et. al., that this is not important for most useful cases.

Notes on Representations

In considering the approximation of the characteristics of the human ear one needs mainly to address the question of frequency scaling and time resolution. Human hearing operates with roughly a logarithmic frequency scale, and most of the speech information is located in the lower part of the range. When speech is analyzed with linear frequency scale techniques, a large fraction of the data points carry very little information, while there may be insufficient resolution around the low frequencies that do contain the information. This appears to be one of the advantages of the wavelet transform, which inherently has a log-frequency scale. An AR-model based spectrum estimate using a log-frequency scale could also meet this requirement.

While the criteria listed above relate to understanding speech representations, they do not form a rigorous framework for comparison. For this reason, a test consisting has been devised which uses each of the representations as the input to a classification system, which is then 'trained' on some real speech data. The resulting classifications are then compared for completeness of coverage of the phonetic classes. Of particular interest here is possible distinctions between the plosive sounds,

which have proven to be much more difficult to classify than the more stable vowel sounds.

THE CLASSIFICATION SYSTEM

The Kohonen Network (Kohonen 1984) is a self-organizing classification system, which performs unsupervized classification of the input patterns into a fixed number of classifications. It is the basis for the experiments performed.

The Kohonen network is based on a two dimensional layer of classic neuron structures, each with complete connection to a single input vector, and with a saturating linear output. The nodes are also connected laterally to all of the other nodes in the network with inhibitory connections, to perform a variation on a winner-takes all function. That is, the node which is most activated by a particular input pattern ends up as the centre of an island of active nodes in the network. The network learns using a simple Hebb style learning rule, with the simplification that the node outputs will be either 1 or 0, and so the active nodes are adjusted in the direction of the input pattern, while those not active are left alone. The system can be efficiently simulated on conventional digital computers or DSP chips.

There are a number of problems with the learning algorithm, which involve the pre-determined rate of learning. This is controlled by a number of parameters which will affect the final convergence of the classifications. This problem has not been solved yet, and for these experiments the learning rate parameters were hand-adjusted until the program worked.

The Kohonen network described only performs a vector quantization operation on the input speech sounds, classifying each time interval into one of a number of categories, so the results presented relate only to instantaneous analysis of the data. Since this is obviously not capable of making classifications on the basis of time sequence information it is not expected to perform particularly well.

A number of researchers have proposed methods of incorporating time information into a recognition process, but time has not allowed any of these more complicated approaches to be tried yet.

EXPERIMENTAL RESULTS

The speech data consisted of two data sets of ten seconds of speech for each of two speakers. The speech was sampled at 10kHz, using the 14bit resolution available on a DASH-16 analog to digital converter card in an IBM-PC clone. This produced four sets of 100,000 samples each. The two speakers were both radio presenters, a female speaker on 4ZZZ-FM, and a male speaker on the ABC. The speech was thus quite fluent and unaffected.

In each speech representation case a window of 64 samples was used (except for the wavelet transform case, where the window varied between 511 samples and 35 samples), with a time resolution of 32 samples. The two AR model based techniques used model orders of 14.

The Kohonen networks were all 18 by 14 rectangular arrays of nodes (not the hexagonal arrangement used by Kohonen), with number of weights per node determined by the input vector length. This was 32 in all cases except the reflection coefficient case, which had only 14 parameters. As per Kohonen's algorithm, the input vectors are length normalized, and the maximally active neuron is determined by minimum Euclidean distance. This proved to be quite effective, although there is some doubt as to whether the information lost in the normalization process is significant, and how this affects periods of silence (which are actually low-level noise).

Unfortunately, it has proved harder to determine 'rate of convergence' and 'completeness of coverage' than anticipated. The experiments performed seem to indicate that all of the test networks *did* converge after the 3000 sample vectors (0.3 seconds), but just where they could be said to have 'converged' is debatable. From the plots generated during training, it seems that the network was largely stable after 1500 sample vectors for the female speaker data set, and after 1000 vectors for the male speaker data set, for *all* representation techniques. This was not expected, so work is continuing with modifications to Kohonen's learning algorithm. Different and more rigorous measurement techniques are also planned.

CONCLUSIONS

We have examined a number of candidate representations of speech, from a loosely theoretical standpoint, with a view to selecting one to form the basis for a speech recognition system using a Kohonen-style neural network classifier. From this discussion of features, the wavelet transform, Zhao-Atlas-Marks (Cone Kernel) or Choi-Williams distributions seem to be the best suited to the speech problem, although this has yet to be shown conclusively in practice.

We have also used four of these techniques (the short-time Fourier transform (STFT), the wavelet transform, auto-regressive (AR) model based spectrum estimation and AR model reflection coefficients), the Wigner-Ville, Choi-Williams and Zhao-Atlas-Marks distributions as inputs to simulations of Kohonen-style networks, and seen that they all converge, for the 10 second dataset used. The rate of convergence seems at this stage to be more dependent on the training data set than on the representation technique used.

Insufficient experimental evidence has accumulated to prefer any particular representation. More work is in progress in this direction.

REFERENCES

Boashash B. (1990) "Time-Frequency Signal Analysis," in *Advances in Spectral Analysis and Array Processing,* (Prentice-Hall Signal Processing Series), S. Haykin, Ed., Prentice-Hall: Englewood Cliffs, NJ), pp 418-517.

Choi H. I. and Williams W. J. (June 1989) "Improved Time-Frequency Representation of Multi-Component Signals using Exponential Kernels," *IEEE Trans. Accoust. Speech and Signal Process.,* 37, no. 6, pp. 862-871.

Cohen L., (July 1989) "Time-Frequency Distributions—A review," *IEEE Proc.*, 77, no. 7, pp 941-981

Daubechies I., (1990) "The Wavelet Transform: a Method for Time-Frequency Localization" in *Advances in Spectral Analysis and Array Processing,* (Prentice-Hall Signal Processing Series), S. Haykin, Ed., Prentice-Hall: Englewood Cliffs, NJ).

Kohonen T. (1984) *Self Organization and Associative Memory,* (Springer-Verlag: New York–Heidelberg–Berlin).

Marple S. L., (1987) *Digital Spectral Analysis with Applications,* (Prentice-Hall: Englewood Cliffs, NJ).

Rabiner L. R. and Gold B., (1975) *Theory and Application of Digital Signal Processing,* (Prentice-Hall: Englewood Cliffs, NJ).

Zhao Y., Atlas L. E. and Marks R. J. (July 1990) "The use of cone-shaped kernels for generalized time-frequency representations of non-stationary signals," *IEEE Trans. Acoust. Speech and Signal Process.,* 38, no. 7, pp 1084-1091.