Transforms for Speech Recognition

Dr. Najmi G. Haider
SZABIST
Karachi, Pakistan

Abstract:
In speech processing applications the microphone acts as a transducer to convert sound to an electrical signal which is further converted to a sequence of discrete samples for digital processing. In this raw form the signal does not show readily discernable useful features, and therefore mathematical transformations have been developed to obtain further information that clearly demonstrates characteristics that can be attributed to various types of sounds comprising speech. This is fundamental to the front-end design of all speech recognizers. The importance of an effective and efficient transform for the speech signal is of prime importance since weaknesses at this foundation stage will undoubtedly impair the performance of the following stages.

This paper discusses transforms for the speech signal with application to Automatic Speech Recognition. It reviews commonly used representations and modifications made to enhance their performance, and other transforms developed for speech processing and recognition.

1. INTRODUCTION

By nature, speech is an acoustic signal which is conveyed to the listener via the medium of air. Sound is conducted through air by variation in air pressure. The ear performs the role of the collector and transducer of sound, and a transformer for the higher levels of processing. In electronic systems the microphone performs the function of converting sound to a continuous electrical signal which undergoes to a further process of conversion to a sequence of discrete time samples for digital processing. In this time domain form the speech signal is not suitable for direct application of recognition algorithms. It has to be subjected to a transformation that will bring out attributes which will more clearly demonstrate attributes that behave in relation to the different sounds comprising speech. This is fundamental to the front-end design of all speech recognizers. The need for an effective and efficient transform for the speech signal is therefore of prime importance since any weakness in the transformation at this foundation stage will undoubtedly impact the overall performance even though the following stages have high performance.

Speech recognizers offer high accuracy rates under controlled conditions but performance is drastically affected from noise under normal environment conditions in which they would be required to operate in practice. A transformation that would enable robust parameters to be derived for representing speech is an area of study that would contribute positively towards the goal of designing speech recognizers that function adequately well in real environments.

The search for a suitable transform should fulfill the following requirements:

i. Enable extraction of information from the acoustical speech signal.
ii. Development and refinement of methods of measurement and observation.
iii. Signal representation and transformation with properties that reveal information more clearly.
iv. There may be a need for masking/suppressing information that may contribute to confusion.
v. Extract a suitable set of features derived from the transformed signal that are uncorrelated and maximize reparability.
vi. Robust to withstand corruption from external sources.

This paper discusses transforms for the speech signal with application to Automatic Speech Recognition. It reviews current established representations, modifications to improvement them, and a new approach, the wavelet transform. Cochlea modeling as an alternative approach is also discussed.

2. TRANSFORMS IN COMMON USE

Several transformations and representations of speech have been employed in the front-end design of speech recognition systems. The choice and suitability of a particular type of representation has been dictated by the complexity of the requirements of the application and the state of technology at the particular time. The most common have been time domain and frequency domain methods.

2.1 Time Domain Methods

2.1.1 Energy, Zero-Crossings, Autocorrelation

The discrete sampled speech signal is a time domain signal and provides elementary parameters and simple methods of processing, using characteristics such as amplitude, energy and zero-crossings. Using these features effective small vocabulary systems have been demonstrated. Autocorrelation methods have been successfully applied for determining speaker pitch frequency characteristics [1].

2.2 Linear Predictive Coding

Linear Predictive Coding (LPC) is a technique [2] that

produced significant improvement in the design and performance of speech recognizers. The vocal tract is modeled as a linear time-varying system. Excitation (air from the lungs) sets it into resonance, the resonant modes being determined by the shape of the vocal tract at the time. When the excitation ceases, these resonances do not stop immediately but continue to 'ring' and gradually wind down, i.e. the future output of the vocal tract is depending on its previous history. A difference equation can be used to predict the future output. This is the basis for applying linear prediction to speech. Although this assumption only holds partially true for speech, it provides an effective method for deriving spectrum characteristics by optimizing the predictor filter parameters. The predictor parameters can also be used to re-synthesize speech. Owing to its simple processing in terms of speed of execution it gained preference over the Fast Fourier Transform approach.

2.3 Frequency Domain Methods

2.3.1 Fourier Transform

The time domain samples are processed and transformed to the frequency domain which shows the characteristics of the speech signal from the view point of frequency components and the respective energy strength. The Discrete Fourier Transform (DFT) represents the sampled signal by a finite set of complex exponentials. This is computationally expensive to process. The Fast Fourier Transform (FFT) is an efficient implementation of the DFT that drastically reduces the time to compute the DFT. Speech is processed using the Short Time Fourier Transform (STFT) technique which provides time resolution as well as frequency resolution [1].

Filter banks in silicon [3] have also been used to derive the frequency spectrum in bands of 16 to 24 channels. The signal is filtered by a set of band-pass filters into a corresponding set of frequency sub-bands where ideally each sub-band would cover a different part of the spectrum and hence collectively would cover the entire frequency range of interest. Since real filters do not have ideal characteristics, the bands of frequency do have some overlap. The speech spectrogram is produced using this method. Digital filter banks have been one of the commonly used methods.

2.3.2 Discrete Cosine Transform

The Discrete Cosine Transform (DCT) [4] is a computationally efficient equivalent of the DFT. It enables an orthogonal transformation of the filterbank channel logarithmic levels producing DCT coefficients that tend to be uncorrelated. In addition, in the case of speech, most of the variance of the original channels is mapped to the lowest three or four DCT coefficients, and so the higher coefficients can be ignored, resulting in a reduction in the dimensionality from the original signal representation enabling it to be processed more efficiently. Thus a 24 channel filterbank output can be represented by 12 coefficients without loss of important information. There are four types of DCT. The forward DCT is used to convert the log magnitude spectrum to cepstral coefficients.

2.3.3 Cepstral Analysis

The speech waveform is considered as the convolution of the excitation and the vocal tract filter impulse response. In the frequency domain this is equivalent to the multiplication of the excitation and filter spectra, which is seen on the short time spectra as a slowly varying envelope which corresponds to the vocal tract filter characteristics, and if the speech is voiced, a rapidly varying fine structure which corresponds to the fundamental frequency of excitation and its harmonics (using a short data window its effects can be reduced). For speech recognition process, only the slowly varying envelope is required. The excitation spectra does not contribute to the recognition process and in fact is a source of confusion, thus the need for a method for separating the two i.e. de-convolution. The best results for doing this have been achieved with the cepstral analysis techniques.

Taking the logarithm of the speech spectra is equivalent to the summation of the logarithm of the excitation spectra and the vocal tract spectra. The process of cepstral analysis is to take the log of the DFT of the speech signal and then perform the Inverse DFT (IDFT) on it producing the Complex Cepstrum. The Real Cepstrum is computed by the logarithm of the magnitude of the DFT followed by the IDFT. The cepstral coefficients so derived define the characteristics of the cepstrum. The pitch period is measured from the point of the first peak in the cepstrum. Deconvolution is achieved by filtering the cepstrum [5]. The low-time cepstrum holds the vocal tract filter impulse response sequence from which the smoothed spectrum can be obtained by performing a DFT on it. The formants can be extracted from this and tracked over time.

3. MODIFICATIONS AND NEW TRANSFORMS

The methods described are the established techniques that are being used in speech processing. In the effort to achieve further improvements at the front-end level, modifications have been made to these methods. Also new methods have been developed to address the inherent weakness in existing methods.

3.1 Frequency Scale Warping

The robustness of the human auditory process has provided the impetus to modify these methods by incorporating features from human auditory perception. As the human auditory sensation has a logarithmic response to tone height, the approach has been to transform the linear frequency domain to a logarithmic one. The lower frequencies are given a finer resolution as

compared to the higher frequencies. Examples of this nonlinear frequency warping are the Bark and Mel scales. Significant improvements in performance have been reported following this modification. The Mel Frequency Cepstrum Coefficients (MFCC) are in predominant use in present ASR front-end designs [6].

3.2 Dynamic Characteristics of Speech

Another characteristic that all methods have in common is the basis of "short time" analysis. The speech signal is analyzed in frames of 10ms to 20 ms on the assumption that owing to the inertia of the human articulatory system, the properties of speech change relatively slowly with time, and the properties of the sound are taken to be fixed over this time frame. This results in the loss of the fine dynamic characteristics of speech which are important in discerning the sound and hence have been ignored. Use of overlapping frames was also not very effective. Thus it was perceived that significant gains could be realized by making use of this information in the speech model.

Adding transitional information such as "velocity" and "acceleration" type parameters has produced further improvements [7]. "Velocity"-type parameters are extracted by taking the difference between successive frames. "Acceleration"-type parameters are the difference in the "velocity"-type parameter between successive frames. Use of MFCC ("velocity") and MFCC ("acceleration") have reduced error rates by 50%.

3.3 Continuous Wavelet Transform

Though the STFT provides time and frequency resolution, its use of a fixed window size limits its resolution capabilities. For the FT to be valid, the signal being processed must be stationary. Speech is a non-stationary signal, but since it is a slowly changing signal, portions of the signal can be assumed to be stationary. Ideally, the width of the window should be equal to the duration the portion of the signal under consideration is stationary for. Therefore, if a signal is stationary for a short duration of time, it will require a short window size to resolve it in frequency and time, and likewise a stationary signal of longer duration would require a longer width window. Clearly, different portions of speech will exhibit stationarity for differing durations, and hence a fixed window size as is used in the STFT is obviously a handicap and will limit its ability to provide good resolution over both frequency and time (the effects of this can be seen on narrow-band and wide-band spectrograms). So it is required for the window size to be adaptive depending on the duration of the stationary portions of speech.

The Continuous Wavelet Transform (CWT) [8] is a new technique developed to address this issue through the implementation of multi-resolution analysis (MRA).

Whereas the FT is the integral over all time of the signal multiplied by a complex exponential function, the CWT is the integral over all time of the signal multiplied by scaled, shifted versions of a wavelet function, called the mother wavelet producing a time-scale view of the signal. Scaling produces compressed or stretched versions of the mother wavelet. Shifting the wavelet is simply the wavelet delayed in time. Thus the CWT process produces wavelet coefficients that are a function of scale and position.

The idea is to analyze different frequencies with different resolutions. In the case of speech, the signal mostly has low frequency components and higher frequency components are less frequent. As seen with STFT, short windows provide good time resolution but poor frequency resolution, while long windows give good frequency resolution but poor time resolution. Therefore the tradeoff is to have good frequency resolution (but poor time resolution) at the lower frequencies (since these frequency components are present most of the time in the speech signal) and good time resolution but poorer frequency resolution for the higher frequencies.

The process to compute the wavelet coefficients is iterative. It is essentially a correlation of the wavelet shifted over the signal. The transformed coefficient value is computed at each shifted position along the signal. A high value indicates presence and strength in the signal of the frequency component represented by the wavelet, and conversely, a low value indicates a lesser influence (or even absence if the value is zero. For the next iteration, the wavelet is scaled (ie. representing another frequency) and the coefficients computed from the correlation of the signal with this new wavelet, using a repeat of the process with the previous wavelet. Processing the signal with all the scaled wavelets completes the picture. A three dimensional graphical display of coefficients vs. scale and time resembles a terrain map. The scalogram is the square magnitude of the CWT and shows the energy distribution of the signal in time-scale plane.

For digital processing, the CWT is discretised by sampling the time-frequency (i.e. scale) plane. The scale axis discretisation is done using a logarithmic scale. The time axis is discretised based on the discretisation of the scale axis. Advantage is taken of the scale change to reduce the sampling rate since as the scale increases; this corresponds to a decrease in frequency hence the sampling rate can also be decreased in accordance with the Nyquist rate, which also implies a reduction in computation.

3.4 Discrete Wavelet Transform

The discretised CWT still requires much computation time as it contains redundant information. The Discrete Wavelet Transform (DWT) offers a considerable reduction in computation time and easier implementation than CWT. It uses subband coding by processing the signal through a series of high-pass filter and low-pass filters. The high-pass filters analyze the high frequencies,

and the low-pass filters analyze the low frequencies. The scale change is performed by upsampling (creating new samples by interpolation between two samples) and downsampling (skipping samples) the signal. Frequency bands that have a significant content in the signal will give high values for the corresponding part in the DWT.

The analysis windows can be chosen to simulate the frequency response of the human cochlea.

4. COCHLEAR MODELS

All the methods described above look at the signal from a digital signal processing viewpoint which is to treat speech as just another signal and to apply DSP techniques in processing it. These are based on the model of "speech production" i.e. speech as a convolution of the excitation and vocal tract filter. Another approach has been to model on the human auditory process from knowledge gained from studies of how the cochlea transforms the speech signal. The ear does not perform a FT or any of the DSP techniques discussed. The auditory nerve fibers are continuously firing (even in the absence of sound) and the incidence of sound affects the rate of firing of these nerves. It seems that phonetic features in speech have a correspondence to the neural discharge pattern, and also shown to be suitable for identifying aspects of speech signal relevant to speech processing and recognition. Auditory models [9] developed along these lines have reported significant results. It is also hoped that these models will also aid in understanding the perceptual properties from the view point of the internal representations of the human auditory system. These auditory based models have also claimed better robustness in noise as compared with the "production" based models.

At present, majority of research in speech processing methods are based on the DSP approach, but Auditory Modeling as an alternative approach is gaining attention as the results have shown promise.

5. CONCLUSION

Transforms are basic to the design of the front-end of speech recognition systems, as weaknesses in the transformation method will surely impair the overall results. Using the established methods, speech recognizers have been able to achieve high recognition rates for large vocabularies by a single speaker. The established methods such as FT, LPC and cepstral analysis are based on the speech production model. Improvements in these models have been obtained by modifications taking account dynamic characteristics of speech and by incorporating perception characteristics of the human auditory system. To address the limitations resulting from uniform frequency resolution of the STFT, wavelet transforms have been applied to speech processing with improved results.

Cochlear Models attempt to reproduce the human auditory system. This is a distinctly different approach which is speech perception-based as opposed to speech production-based, the obvious motivation for this being the performance of the human auditory system.

The critical issue in implementation of speech recognition systems is noise. Besides the techniques discussed, there are several mathematical transforms in existence. These transforms could be studied from the view point of robustness to external disturbance factors, i.e. inherent properties of these transforms that make them resistant to noise.

Another issue is speaker independence for which transforms could be studied to resolve also.

REFERENCES

[1] L.R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals". Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1978).

[2] J. Makhoul, "Linear Prediction: Atutorial Review," Proc. IEEE, Vol. 63, pp. 561-580, 1975.

[3] Andrew R.B., Gabriel M.R., "Micromachined Micropackaged Filter Banks", IEEE Microwave and Guided wave Letters, Vol. 8, No. 4, April 1998.

[4] V. Sanchez, P. Garcia, A.M. Peinado, J.C. Segura, A.J. Rubio, "Diagonalizing properties of the discrete cosine transforms" IEEE Trans. Signal Processing, Vol. 43, pp2631-2641, November 1995.

[5] A.V. Oppenheim, R.W. Schafer, "Homomorphic Analysis of Speech", IEEE Trans. On Audio and Electroacoustics, Vol. AU-16, No.2, pp. 221-226, June 1968.

[6] L. Jia, B. Xu, "Including Detailed Information Feature in MFCC For Large Vocabulary Contious Speech Recornition", ICASSP 2002, Vol. 1, pp 805-808.

[7] S. Furui, "Speaker Independent Isolated Word Recognition using Dynamic Features of the Speech Spectrum", IEEE Trans. on ASSP, Vol. 34 (1) 1986, pp 52-59.

[8] O.Rioul and M.Vetterli**,** "Wavelets and signal processing", IEEE Signal Processing Magazine, Vol.8, No. 4, Oct. 1991, pp.14-38.

[9] R.F. Lyon, "Computational Models of Neural Auditory Processing", Proc. ICASSP (1984) 36.1.1-36.1.4.