An Introduction to Front-End Processing and Acoustic Features for Automatic Speech Recognition

Magnus Rosell

January 16, 2006

Abstract

This is a term paper in the course Speech Technology within the Swedish national graduate school of language technology, GSLT. It gives an introduction to front-end processing and acoustic features for automatic speech recognition. That is, it describes the process of transforming the physical sound signal of speech into a feature vector that is suitable for further processing. The process is divided into a few steps that are described separately. For some of the steps different approaches are described.

1 Introduction

This is an introduction to front-end processing and acoustic features for automatic speech recognition (ASR). The front-end of an ASR system is the part that transforms speech to a vector of features that is suitable for further processing. The vector can contain many different kinds of features, like face-expressions, gestures etc.

Here we look at the features that may be derived from the physical sound signal, the acoustic features. This paper is based mainly on (Holmes and Holmes, 2002), (Picone, 1993), and the Internet tutorials (Robinson, 1998) and (Ibarra and Curatelli, 2000).

The aims behind the front-end processing are:

- The parameters/features should capture the salient aspects of the speech signal. These should also be perceptually meaningful if it is possible. To capture the salient aspects it is necessary to capture the spectral dynamics, the change of the spectra over time.
- The features should be robust in the sense that the particular task should not be affected by the distortions that can appear, due to among other things enviromental aspects and/or transmission medium. For instance should a general ASR-application be able to recognize speech from different persons.

Figure 1, from (Picone, 1993), divides the digital signal processing (DSP) of the front-end into three parts: spectral shaping, spectral analysis, and parametric transform. Statistical modeling is sometimes thought of as a part of the speech recognizer. The features or parameters that are a result of the front-end are analyzed statistically to define similarity between feature vectors and sometimes also to reduce the size of the representation (eg quantization). Statistical modeling will not be discussed further.

Spectral shaping is the process of converting the anlogoue sound signal into a digital signal, A/D conversion. It also often involves some filtering, emphasizing important frequency components.

Spectral analysis is what the name suggests – the analysis of the spectrum in order to capture the salient aspects of the signal.

Parameter transform is the name for the moulding of the measurements achieved through spectral analysis. Here the different features are put together in one vector. At this stage the differences between consecutive frames (small time intervals) can be used to capture the dynamics of the signal.

Channel normalization is not accounted for in the picture. It is a process in which the properties of the medium is taken into consideration. By the medium we mean the environment in which the speech was uttered and the channel through which it reached the spectral analyzer. Channel normalization is applied after the spectral analysis and is often accomplished by analyzing a short time interval, so we consider it a part of the parameter transform.

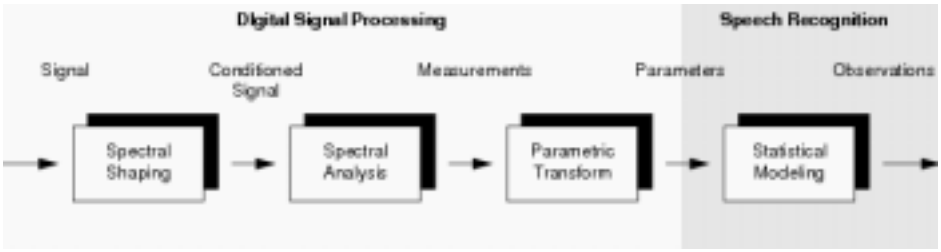


Figure 1: Digital Signal Processing (from (Picone, 1993))

2 Spectral Shaping

To convert an analogue sound wave to digital form one first filters the signal (reduces frequency information) and then samples it. Theoretically, to represent a frequency of xHz a sampling frequency of $2x$Hz is needed. The ear can perceive a dynamic range (amplitude ratio) of 20 bits (1 to 10^6). Normally each sample is stored in 16 bits, for telephone speech 8-12 bits are used.

The intensity of the speech sound is not evenly distributed over the frequency range. It falls with approximately 6 dB per octave. Often the signal is pre-emphasized to compensate for this.

3 Spectral Analysis

To analyze the spectrum the sampled signal is divided into frames, short time intervals. If these frames are short enough the signal appears static and can be represented by a feature vector. On the other hand, they have to be long enough to contain at least one cycle of the lowest frequency one is interested in. A normal length is 20-25 ms, which includes a few periods of the glottis. Usually the frames are set to overlap so that their centers lie only 10 ms apart. Each frame is multiplied by a tapered window, so that the values near the edges become zero. This prevents the discontinuities at the edges to affect the result of the further processing.

The sound signal is represented by a feature vector for each frame. The methods for extracting features for a single frame, described in the following sections, could be combined in several ways. Figure 2 shows some of the possibilities.

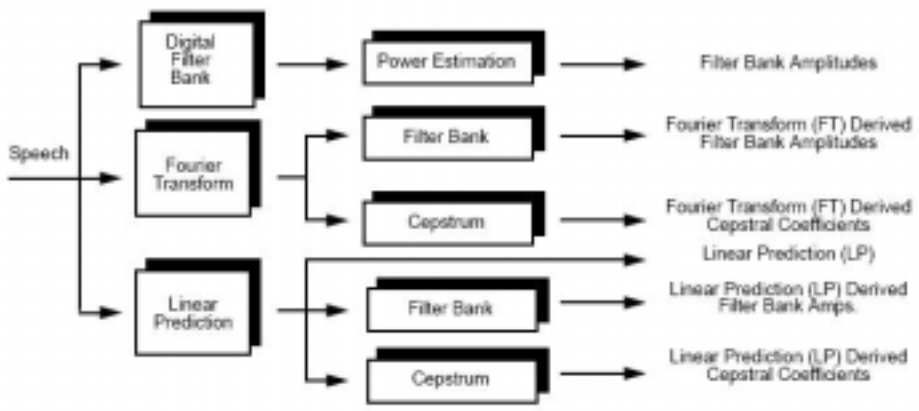


Figure 2: Spectral analysis (from (Picone, 1993))

3.1 Filter Banks

The information needed in each frame is a description of the frequency distribution, i.e. how the power of the signal is distributed over different frequencies. A filter bank, in its simplest form, is a set of bandpass filters with different frequencies covering the interesting part of the spectrum (the fundamental frequency and the formants). The output of the filters during a frame can be used as features.

The center frequencies of the filters can be chosen in several ways. Usually they are set according to some perceptually motivated scale. The perceived pitch of a sound is not equal to the actual frequency. A popular approximation of the real mapping is the mel scale, the mel frequency (f_{mel}):

$$f_{mel} = 2595 \log_{10}(1 + f/700.0), \tag{1}$$

where f is the actual frequency. An increase in frequency is easier to recognize in the lower register than in the higher.

The human auditory system can not distinguish between frequencies that are close to each other. The higher the frequency the bigger are these critical bands (the intervals within which the frequencies can not be separated from the center frequency). Using for instance the mel frequency the size of the critical band is approximated by:

$$BW_{\text{critical}} = 25 + 75[1 + 1.4(f_{mel}/1000)^2]^{0.69}. \quad (2)$$

In a critical band filter bank the bandpass filters are linearly spaced on a perceptually motivated scale (for instance the mel scale). The bandwidth of the filters are set to the critical bandwidth for the center frequency.

An important approximation to a critical band filter bank is based on the mel scale. Between 100Hz and 1kHz ten filters are spaced linearly. Above 1kHz five filters are assigned for each doubling of the frequency. That is, they are logarithmically spaced. The bandwidths are set so that the 3dB point (half the power) is half-way between the centers of consecutive filters. Normally samples from the 20 first filters are used.

Figure 3 shows the critical band filter bank and mel-filter bank, including the pre-emphasis usually connected to them.

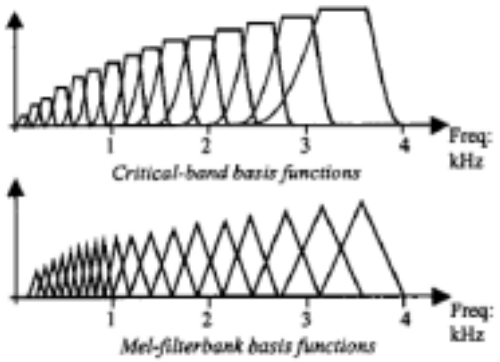


Figure 3: Two perceptually motivated filter banks (from (Milner, 2002))

Another way to a accomplish a corresponding feature vector is to use the Fourier transform to sample the the signal at the desired frequencies (like for instance the centers of the filters in the mel-filter bank). This gives a Fourier filter bank and is accomplished using the Discrete Fourier Transform (DFT)[1]:

$$S(f) = \sum_{n=0}^{N_s-1} s(n)e^{-i\frac{2\pi f}{f_s}n}, \quad (3)$$

where f is the frequency, $S(f)$ is the Fourier coefficient for the particular frequency, $s(n)$ is the sampled speech signal, f_s is the sampling frequency and N_s is the number of samples in the window under consideration. Often the signal is sampled at a higher resolution in frequency and then the values for each "filter" is calculated as an average over several frequencies.

[1]When appropriate one use the Discrete Cosine Transform (DCT) or the Fast Fourier Transform (FFT) rather than the DFT.

3.2 Linear Prediction

In linear prediction (LP) the signal $s(n)$ is modeled as a linear combination of the previous samples:

$$s(n) = \hat{s}(n) + e(n) = \sum_{i=1}^{N_{LP}} a_{LP}(i)s(n-i) + e(n), \tag{4}$$

where $\hat{s}(n)$ is the model, $a_{LP}(i)$ are the coefficients that need to be decided, N_{LP} is the order of the the predictor, i.e. the number of of coefficients in the model, and $e(n)$ is the model error, the residual. There exit several methods for calculating the coefficients. The coefficients of the model that approximates the signal within the analysis window (the frame) may be used as features, but usually further processing is applied (see below).

The higher order that is used the better the model predicts the signal. A lower order model, on the other hand, captures the trend of the signal, ideally the formants. This gives a smoothed spectrum.

The LP coefficients give uniform weighting to the whole spectrum, which is not consistent with the human auditor system. Perceptual linear prediction (PLP), introduced in (Hermansky, 1990), incorporates a physiologically motivated weighting of the frequencies.

The left side of Figure 4 shows PLP. Before the linear prediction - the auto-regression (AR) modeling - a few steps that simulate the auditory system are taken. The signal is Fourier transformed (spectral analysis) and mapped to a physiologically motivated frequency scale (critical-band analysis). Then the unequal sensitivity of human hearing across frequency is compensated for by pre-emphasis (equal-loudness pre-emphasis). Last the non-linear relation between intensity and perceived loudness is modeled by taking the cubic root of the intensity (intensity-loudness power law). The last step in the picture is discussed in the next section.

3.3 Cepstral Analysis

The spectrum of a speech signal is the result of the sound source (in voiced speech the oscillating vocal folds) and the vocal tract acting like a filter. The fundamental frequency of the glottis is resonated resulting in harmonics and partials, some of which are strengthened and others weakened by the vocal tract. In the time domain (the ordinary sampled speech signal) the signal $s(n)$ is modeled by a convolution:

$$s(n) = g(n) \otimes v(n), \tag{5}$$

where $g(n)$ is the excitation signal, the signal coming from the glottis, and $v(n)$ is the vocal tract filter. In the frequency domain (after for instance a Fourier transform) this is a product:

$$S(f) = G(f)V(f). \tag{6}$$

By instead considering the logarithm of the intensity the filtering process becomes a simple addition:

$$\log(S(f)) = \log(G(f)) + \log(V(f)). \tag{7}$$

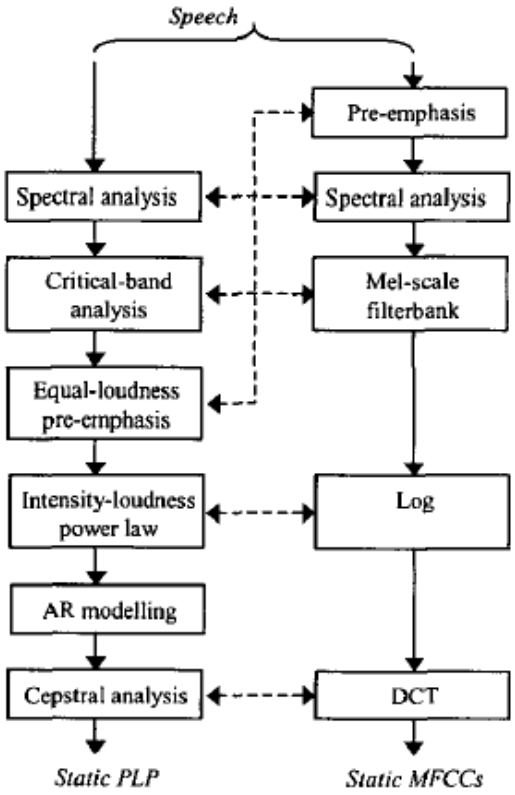


Figure 4: PLP vs MFCC (from (Milner, 2002))

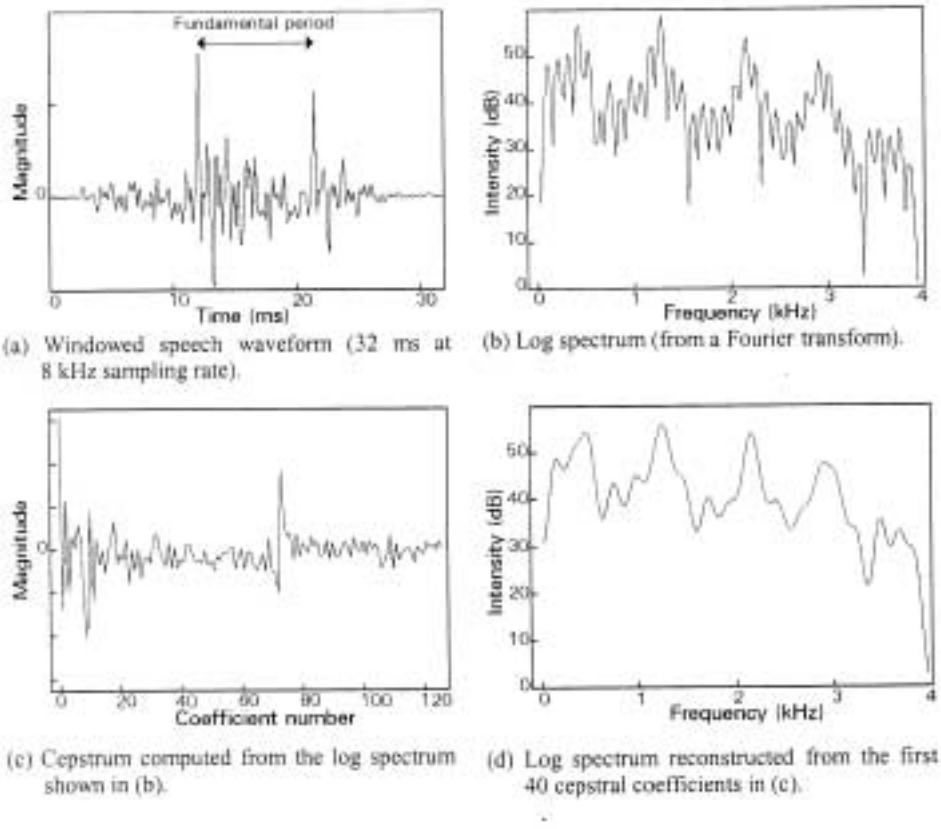
Hence the logarithm of the spectrum can be viewed as a superposition of the source or excitation signal and the vocal tract resonance.

Figure 5 shows the steps of cepstral analysis. The spectrum of a sound signal may also be viewed as a superposition of several waves in the frequency domain. The waves with high frequency correspond to the harmonics of the excitation signal, while the more slowly varying represent the vocal tract shape. The cepstrum is the series of coefficients when doing the Fourier transform of the spectrum. In this new domain the lower order coefficients represents the slowly varying parts of the spectrum. Usually there is a spike in this series that corresponds to the harmonic series of the vocal folds. Considering only the coefficients lower than this spike gives a representation of the vocal tract shape.

By truncating the cepstral coefficient series and using the inverse transform a smoothed spectrum is achieved. This is called a cepstral smoothing and ideally shows the slow trends in the spectrum, i.e. the formants, see Figure 5.

The perhaps most common use of cepstral analysis is when constructing the mel-frequency cepstral coefficients (MFCCs). The right side of Figure 4 shows the steps involved.

As can be seen in Figure 4 cepstral analysis is often also applied to the output of a linear predictor, since it has proven beneficial.



(a) Windowed speech waveform (32 ms at 8 kHz sampling rate).

(b) Log spectrum (from a Fourier transform).

(c) Cepstrum computed from the log spectrum shown in (b).

(d) Log spectrum reconstructed from the first 40 cepstral coefficients in (c).

Figure 5: Cepstral Analysis (from (Holmes and Holmes, 2002))

3.4 Energy and Pitch

Apart from the frequency information that is discussed above, the energy or intensity of the signal is included in the representation in some way.

It is difficult to reliably estimate the fundamental frequency from the speech signal, so it is often neglected, though it may be benificial to include it. One way of extracting it is thorough cepstral analysis (the spike in the coefficients, see above).

4 Parameter Transform

The spectral analysis results in feature vectors for each frame. These represent the actual speech as mediated through the environment in which it was uttered and through the A/D-conversion and transmission (i.e. transmission through for instance a telephone line).

4.1 Channel Normalization

Channel normalization is the process of compensating for distortions due to the mediation of the signal. Distortions could be caused by the environment (other sounds in the background and/or acoustics of the room) and during transmission.

There are two main methods for reducing the distortion. In cepstral mean normalization (CMN) or subtraction (CMS) (Furui, 1981) the average of the cepstral coefficient feature vectors over some interval is calculated. Every feature vector is then subtracted by this average vector. This reduces the impact of stationary and slowly time-varying distortion.

RASTA (relative spectral) filtration (Hermansky and Morgan, 1994) use a band-pass filter on the time series of each coefficient from the spectral analysis. Figure 6 shows the idea. Each coefficient is filtered on its own, i.e. it is possible to have different settings for them. The compressing static nonlinearities may be the logarithm of the signal, and when that is the case the expanding side is the exponential function. The band-pass filter removes the fast and slow changes of the signal relative to a time interval. The aim is to keep only the changes that are due to the speech sound.

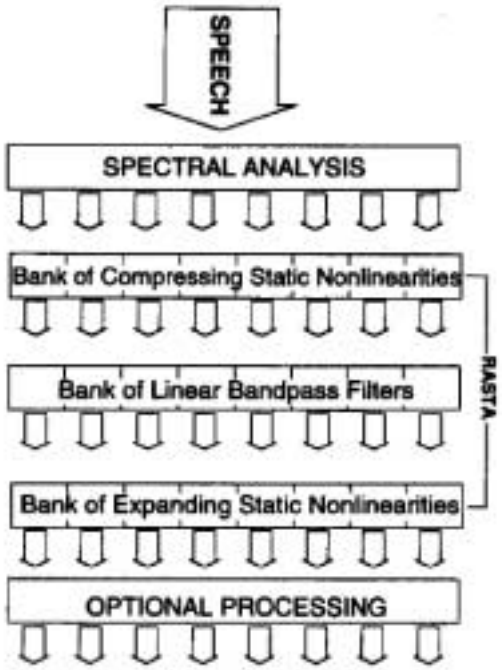


Figure 6: RASTA filtering (from (Hermansky and Morgan, 1994))

4.2 Dynamic features

The features described so far capture the average frequency distribution during a frame. Much of the information in the speech signal is however in the change of it, in its dynamics. To capture this often the time derivative is approximated by differencing between frames after and before the current, for instance:

$$\Delta y_i = y_{i+d} - y_{i-d}, \tag{8}$$

where y_i is the feature vector at frame i, and d typically is set to 1 or 2. To avoid problems due to noise and other random fluctuations in the signal

linear regression can be applied to the same set of feature vectors. Adding the Δy_i:s to the feature vector usually results in better performance.

By differencing between the approximations of the time derivative (or using linear regression again), the second order derivative can be approximated. Adding these to the feature vector also often improve performance, but not as much as the first order derivatives.

5 Comparing Spectral Analysis Methods

This section contains a brief comparison of some of the different approaches to spectral analysis (SA) outlined above.

5.1 MFCC vs LP

In (Davis and Mermelstein, 1990) a few SA methods are compared in a syllable-oriented speaker dependent speech recognition system. The experiments were made in a noise-free environment and the segmentation was done manually. They found that features derived using cepstrum analysis outperform those that does not use it and that filter bank methods outperform LP methods (PLP methods were not included). Best performance was achieved using MFCCs.

5.2 MFCC vs PLP

The use of PLP and MFCC give comparable results according to (Holmes and Holmes, 2002).

5.2.1 In theory

A theoretical comparison of MFCC and PLP analysis is given in (Milner, 2002). Figure 4 shows their perspective. The broken arrows connect the processing stages that are similar. Both methods use a Fourier transform on similarly windowed speech to do the spectral analysis. Figure 3 shows the critical band analysis weighted by pre-emphasizing of both methods. These are very similar in principle. The intensity-loudness power law of the PLP analysis corresponds to the logarithmic compression in the MFCC analysis.

It is when the AR modeling comes in that the two methods diverge. In MFCC analysis the log mel-filter bank is transformed using the Discrete Cosine Transform (DCT). In PLP the output of the critical band filter bank is approximated using the first few coefficients in a LP model. This approximation is then transformed and the first cepstral coefficients are used as features for the signal.

5.2.2 In practice

The theoretical comparison in (Milner, 2002) continues with a practical implementation. The spectral analysis is followed by channel normalization (both RASTA and CMN are tried) and extraction of dynamic features. The best results are reported for MFCC with RASTA filtration.

6 Conclusion

The front-end of an ASR system extracts features from the physical speech signal. These are put together in a feature vector that is used in the rest of the system. There exist several methods that can be used to accomplish this, some of which have been discussed.

Section 5 only gives a brief comparison of methods. My guess is that what is the best method has to be investigated for each particular setting.

References

Steven B. Davis and Paul Mermelstein. 1990. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *Readings in speech recognition*, pages 65–74, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. ISBN 1-55860-124-4.

S. Furui. 1981. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Signal Processing*, 29(2):254–272.

H. Hermansky. 1990. Perceptual linear predictive (plp) analysis of speech. *Journal of Acoust. Soc. Am.*, 87(4):1738–1752.

H. Hermansky and N. Morgan. 1994. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589.

John Holmes and Wendy Holmes. 2002. *Speech Synthesis and Recognition*. Taylor & Francis, Inc., Bristol, PA, USA. ISBN 0748408576.

Oscar Mayora Ibarra and Francesco Curatelli. 2000. A brief introduction to speech analysis and recognition, an internet tutorial. URL `http://www.mor.itesm.mx/ omayora/Tutorial/tutorial.html`.

B. Milner. 2002. A comparison of front-end configurations for robust speechrecognition. *ICASSP '2002*, 1:797–800.

J. W. Picone. 1993. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):119–1215.

Tony Robinson. 1998. Speech analysis. URL `http://svr-www.eng.cam.ac.uk/~ajr/SpeechAnalysis/SpeechAnalysis.html`.