EFFECT OF NOISE AND DISTORTION IN SPEECH ON PARAMETRIC EXTRACTION

B. Yegnanarayana
Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore 560012, INDIA

Summary

Parameter or feature extraction from speech signal forms the basis for systems designed for speech recognition, speaker verification, speech bandwidth compression etc. The parameters in general are critically dependent upon the short-time spectrum of speech. The input speech waveform is however, subjected to several types of noises and distortions due to background noise sources, reverberation, close speaking into a microphone, telephone system imperfections etc. These factors modify the spectrum of the speech signal and hence the parameters extracted.

Characteristics of common sources of noise and distortion are described in this paper and their effect in shaping the spectrum of speech is discussed. Steps to reduce the influence of some noises while producing speech input to a system are suggested. Methods of normalization of spectral distortions due to noise and the effect of such normalization on parametric extraction are also discussed.

Introduction

Digital processing of speech signal is performed to extract features for speech recognition and verification systems or to obtain parameters to represent speech information for bandwidth compression systems. Some of the features of interest are formants of the vocal tract, fundamental frequency (pitch) of voiced speech, poles and zeros of the vocal tract system and glottal pulse shape. The parameter sets that are generally used to represent speech information are autocorrelation coefficients, cepstral coefficients, linear predictor coefficients and formants. The linear predictor coefficients are being extensively used in view of the simplicity of their computation directly from the speech wave[1].

In parametric representation of speech the parameter set is obtained so as to approximate the envelope of short-time spectrum of a speech segment as closely as possible. For example the linear predictor coefficients (LPC) are derived by minimizing the integrated ratio of the actual spectrum of the speech signal and the spectrum of the assumed all-pole model[2]. After minimization the model spectrum approximates the speech spectrum mainly at its peaks. Some of these predictor coefficients can be used directly as features for speaker verification[3]. Deviation of the parameters from their true values can lead to misrepresentation of the features. As the parameters are critically dependent upon the shape of the short-time spectrum any change in the spectral shape is automatically reflected in the values of the parameters.

Spectral distortions in speech waveform can occur due to several reasons. Firstly the environment in which the speech is produced contributes to the changes in speech spectrum. For example, in a computer room or in a general office room there will be additive background noise which could be either steady or impulsive, and multiplicative (in spectral domain) reverberant noise. Secondly the electroacoustic system used for transducing the speech will impose its frequency response, thus modifying the spectrum of the signal. Finally the pre-processing adopted before digitizing speech and the choice of window width will also influence the parameters extracted from the speech signal. It is the object of this paper to study the characteristics of various types of sources that affect the features or parameters extracted from speech signal. We shall discuss the way the short-time spectrum of speech is altered under different conditions of recording and suggest suitable methods for correcting the spectrum. We shall also discuss the steps to be taken while producing speech input to a signal processing system to overcome the problems arising out of some of the common sources of noise.

Sources of Noise and Distortion

The characteristics of different types of noise and distortion which contribute to changes in the shape of the speech spectrum are discussed in this section.

Additive Background Noise

The background noise in a normal live room may be produced by airconditioning units, fans, flourescent lamps, typewriters, computer system, conversation among people etc. Typical noise levels of various units in a computer installation are given in Table 1.

TABLE 1: Noise Levels (dB) in a Computer Room

Source \ Weighting network	A	B	C
Overall level	68	70	73
Line Printer	81	81	81
Card Reader	78	78	78

Generally most of these noises have continuous frequency spectra[4] and are additive in nature. In some cases the spectra may have significant energy in some frequency bands as for example the noise from a ringing telephone where the energy is concentrated mostly in the band 1200-2400 Hz. It should be noted that the frequency spectra reported in literature for various types of noises are long time average levels whereas the short-time spectrum of speech is affected by the noise component within the analysis interval. To illustrate the effect of noise on parametric extraction, LP coefficients obtained for a speech segment under different simulated conditions of noise are given in Table-2. The noisy speech signal is generated by adding samples of band-limited (0-3.5 KHz) gaussian noise to samples of a voiced speech segment. The sampling rate was 8 KHz and a 12th order predictor was used. Only the first six coefficients are listed in the table.

Room Reverberation

If the room in which speech is produced has hard reflecting surfaces, then there will be a significant component of reverberant sound along with the direct sound. The total mean squared acoustic pressure at a point distance 'r' from the speaker inside a live room is equal to sum of the mean squared pressure due to direct field which is proportional to $1/4\pi r^2$ and the mean square pressure due to reverberant field which is proportional to $4/a$ where 'a' is the total acoustic absorption present inside the room. For a given total absorption the direct field predominates over reverberant field only if the distance of the speaker from the microphone is less than $\sqrt{16\pi/a}$.

On the other hand, if there are only a few discrete reflections the speech spectrum $S(\omega)$ is modified as

$$S(\omega) \sum_{k=1}^{N} A_k e^{-j\omega\tau_k}, \text{ where } A_0(=1) \text{ is}$$

the direct component at the microphone and A_k, k = 1,2,.. N are the contributions of the discrete reflections at times τ_k, k=1, 2,... N. The spectrum of speech signal is thus multiplied by a function which has periodic components (τ_k) in the frequency domain. In Table 2 the effect of discrete reflections on LP coefficients is also illustrated.

Distortion due to Close Speaking into a Microphone

While speaking very close to a microphone the variation of distance and relative orientation between the talker and the microphone can cause significant changes in the level of the speech signal. This effect is equivalent to a slow variation of the gain of the electro-acoustic system with time. Table 3 gives the average speech intensity at different distances of talker from microphone when the talker is producing speech at normal conversational level. It is evident from the table that close speaking can produce large fluctuations in the speech level even for a slight movement of the talker. This type of distortion is particularly significant for speech input into a telephone where the average distance is only about 2".

TABLE 3: Normal Conversational Level (in dB) at Different Distances from Microphone

Distance \ Weighting network	A	B	C
0 - 1/2"	92	96	98
2"	88	92	95
4"	83	87	90
6"	78	80	84
9"	74	78	81
12"	72	74	76
18"	71	72	74

Frequency Response of a Telephone System

The frequency response of a carbon microphone used in a telephone has a significant peak in the frequency range 1500-2500 Hz and several minor peaks beyond 2500 Hz. This response will be superimposed on the speech signal and may result in producing significant

TABLE 2: Effect of noise and discrete reflections on LP coefficients

Signal \ LPC	a_1	a_2	a_3	a_4	a_5	a_6
a) Speech samples	-2.701	3.845	-3.786	2.961	-1.744	0.591
b) Speech + Noise SNR = 14 dB	-1.031	0.208	0.138	0.182	-0.009	-0.202
c) Speech + Noise SNR = 8 dB	-0.668	-0.032	0.045	0.237	0.107	-0.167
d) Discrete reflections $A_1 = 7$, $A_2 = .6$ $\tau_1 = 30$, $\tau_2 = 50$	-2.900	4.647	-5.388	5.185	-4.095	2.587

variations in the values of the parameters extracted.

Prefiltering and Aliasing

Speech input to a digital processing equipment is prefiltered using a suitable low-pass filter and then sampled at a convenient rate. Sampling is performed at a sufficiently high rate, usually greater than twice the cut-off frequency of the low-pass filter, to reduce the aliasing errors in the spectrum. However, the frequency characteristics of the filter including its roll-off near the cut-off frequency are superimposed on the spectrum of the speech signal. In the frequency range of analysis (i.e., half the sampling frequency) the parameters derived to approximate the spectrum try to account for the sharp roll-off also. In LP analysis this increases problems of ill-conditioning of the autocorrelation matrix.

Normalization of Noise

In this section we shall consider methods available to reduce the effect of noise and distortion on parametric extraction.

Additive Noise

Let us first consider the case where the signal and noise are additive i.e., the recorded speech $x(t)$ is

$$x(t) = s(t) + n(t)$$

where $s(t)$ and $n(t)$ are the speech signal and noise respectively. An obvious way to reduce the effect of noise is to increase the speech level relative to the noise level. But in a high background noise environment, as in a computer room, it is not always possible to maintain high signal to noise ratio. Assuming the noise to be stationary and uncorrelated with the speech signal, the short-time spectrum of $x(t)$ is obtained as

$$X_t(\omega) = S_t(\omega) + N_t(\omega)$$

where the subscript 't' indicated that the spectrum is a function of time also. $N_t(\omega)$ is the spectrum of the sample function of the noise present in $X_t(\omega)$ and is in general not equal to $N(\omega)$, the stationary spectrum of the noise. The problem is how to reduce the effect $N_t(\omega)$ knowing $N(\omega)$. Four possibilities may be considered for discussion.

1. Subtract $N(\omega)$ from $X_t(\omega)$: This does not yield a satisfactory result since $N(\omega)$ is the long-time average spectrum whereas $N_t(\omega)$ is the spectrum of the noise present only in the analysis frame. $N_t(\omega)$ will usually have random fluctuations and hence, for the same overall noise level the parameters extracted from a given speech segment would be different for different sample functions of the noise.

2. Whitening the Noise Component: The noisy speech signal $x(t)$ is passed through a filter whose frequency response is given by $\sqrt{1/N(\omega)}$. The output signal spectrum is

$$\left[S_t(\omega) + N_t(\omega)\right] / N(\omega)$$

Even if we assume that this process whitens $N_t(\omega)$ portion approximately, the spectrum of the signal component i.e., $S_t(\omega)/N(\omega)$ is not obviously equal to $S_t(\omega)$. Thus the signal spectrum is also modified by $N(\omega)$. In particular, if $N(\omega)$ has several significant valleys they will appear as peaks in the resulting speech spectrum. Thus subtraction of log spectrum of the noise from the log spectrum of the noisy speech signal may not provide a satisfactory solution.

3. Feature Extraction from Selected Portions of the Spectrum: If $N(\omega)$ has large amplitude in certain frequency bands, then signal to noise ratio is likely to be low in those regions. One of the ways of extracting reliable features is to use only those frequency regions of the spectrum where the SNR is known to be large and ignore the low SNR regions. A particularly convenient method of implementing this scheme is by means of selective linear prediction analysis proposed by Makhoul[5].

4. Autocorrelation Method: Normalization of long-time spectrum of noise can be performed by using a second-order inverse filter derived from the average values of the first two autocorrelation coefficients[6]. The filter normalizes only the gross spectral distributions of a speech utterance.

Reverberation

Reverberation is a multiplicative noise and cannot easily be suppressed. While the autocorrelation function of speech in the presence of additive noise can be used to extract certain features such as pitch, similar processing cannot be applied for reverberant speech. The autocorrelation of noise is usually small for large intervals. Also the effect of additive noise can be reduced by increasing the signal level. On the other hand reverberation is unaffected by the signal level. A straightforward method is to reduce the reverberation time of the room by acoustic treatment. But this may not always be possible. The only other alternative available is to increase the level of direct sound at the microphone relative to reverberant level. This can be achieved by speaking very close to the microphone.

Close Speaking into a Microphone

As explained earlier, close speaking into a microphone is likely to result in changes in output level of electroacoustic system due to movements of the speaker relative to the microphone, and also due to directional characteristics of the microphone and the radiation from mouth. This effect can be reduced by maintaining a constant distance from the microphone or by

providing an automatic gain control in the amplifier to compensate for gross variations in level. Distortions caused by variations in level have negligible effect if feature extraction is confined to a short segment (20-40 msec) of speech during which the level may be assumed to be constant.

It is generally preferable to speak at a distance of about 9" from microphone as this will produce insignificant variation in levels due to movement of speaker's head. From Table-3 it is evident that at 9" a variation of ± 3" in distance can produce a change in level of about ± 3 dB whereas at 4" even a variation of ± 2" can result in level changes upto ± 6dB. But the disadvantage of speaking at an average distance of 9" is that the average speech level will be only about 80 dB in contrast to 90 dB at 4" which is desirable to overcome background noise and reverberation effects.

Conclusions

Noise from environment severely limits the performance of a speech signal processing system. Various sources of noise and distortion have been identified and the nature of distortion they produce is discussed. Quantitative information on the effect of these distortion on parametric extraction is being obtained. At present normalization of noise can be done only in a very limited number of situations. Until satisfactory signal processing techniques are developed to remove the effects of noise, the best solution is to provide a high SNR conditions for inputting speech to a signal processing system.

Acknowledgement

The author wishes to thank Prof. D. Raj Reddy, Computer Science Department, Carnigie, Mellon University, USA, for initially suggesting the problem and for the interesting discussions with him on this subject.

References

1. B.S. Atal and S.L. Hanauer, 'Speech analysis and synthesis by linear prediction of speech wave', J. Acoust. Soc. Amer., Vol. 50, No.2, pp. 637-655, 1971.

2. J. Makhoul, 'Linear prediction: A tutorial review', Proc. IEEE, Vol. 63, pp. 561-580, April 1975.

3. A.E. Rosenberg and M.R. Sambur, 'New techniques for automatic speakers verification', IEEE Trans. Acoust., Speech and Sig. Processing, Vol. ASSP-23, pp. 169-176, April 1975.

4. L.L. Beranek, 'Noise Reduction', New York; McGraw Hill, 1960, Part 4.

5. J. Makhoul, 'Spectral linear prediction: properties and applications', IEEE Trans. Acoust., Speech and Sig. Processing, Vol. ASSP-23, pp. 283-296, June 1975.

6. F. Itakura, 'Minimum prediction residual principal applied to speech recognition', IEEE Trans. Acoust., Speech, and Sig. Processing, Vol. ASSP-23, pp. 67-72, Feb. 1975.