THE LONG-TERM SPECTRUM OF VOICE

Jeffery Pittam * and J. Bruce Millar **

* Brisbane College of Advanced Education
** Australian National University, Computer Sciences Lab.

ABSTRACT - This paper presents an analysis of published information about the long-term spectrum of the voice. The historical development of the measure is first examined leading to a classification of the published works. Techniques used to compute the LTS are then presented, and the utility of the spectrum to various applications is considered. The outcome of this work is a research tool in the form of an annotated and classified bibliographic database.

INTRODUCTION

This paper reports a multi-disciplinary analysis and classification of published information about the long-term spectrum (LTS) of the human voice. This reputably stable measure has generated interest in a wide range of application areas where there has been a need for a global measure of the speech of an individual or of a representative group of speakers. The basic appeal of the LTS is that it aims to describe the spectral characteristics of speech as a whole by averaging the contribution of individual speech sounds. It is a simple measure in that it ignores the temporal structure of speech and can be applied to any substantial portion of speech without regard for the precise timing of its onset or duration. It has been found by several researchers working with different language material that the LTS tends to stabilise in shape after averaging approximately 30 seconds of continuous speech. That is, all other things being equal, the shape of the LTS will not change significantly no matter how much more speech is added.

The report examines LTS from two perspectives. Firstly, the historical development of the LTS is presented, showing major areas of application and measurement. This leads to the introduction of the classification system we have adopted. Our current bibliography contains over 180 items dating from 1917 to 1988. Secondly, we examine the quantitative methods used to measure the LTS and the utility of the spectrum to specific applications. Included in this section is a critical analysis of the range of analytic techniques that have been used to compute the LTS measurement. These have developed from bandpass filtering, through Fourier analysis, and include other formulations such as linear prediction analysis, partial correlation analysis and cross correlation. In addition, a number of measures have been derived from the LTS, and a range of multi-variate statistical techniques have been applied to the set of LTS data points.

HISTORICAL DEVELOPMENT

A graph of the energy distribution of speech had been produced as early as 1917 (Crandall, 1917). This was not a long-term spectrum as such, however, being a synthetic spectrum based on the energy distribution in English vowels which were weighted for frequency of occurrence. By 1922 a test sentence had been recorded and analysed (Crandall & Mackenzie, 1922), but the resulting spectrum was still not a measure of continuous speech. On this occasion, each syllable of the test sentence was said slowly and the energy levels noted. From this a mean was calculated. The spectrum actually presented was a composite of the mean spectrum of six speakers - four male and two female. The first report of the spectral analysis of continuous speech seems to have been Sivian (1929).

These early studies were conducted on 'normal' voice and speech, and were aimed specifically at gaining the knowledge necessary for the development of telecommunications systems. This remained much the same until after the second world war. During World War II, however, new applications arose as researchers studied the effect of high altitudes on voice and speech, and how gas and oxygen masks distorted speech (Stevens, Egan & Miller, 1947).

These first 30 years of the measure's development saw fewer than 20 studies published in which it was used, and all but two of these were conducted in the USA. Since 1950, the situation has

changed dramatically, with researchers in some 16 countries publishing studies that use the LTS. The areas of interest have also expanded considerably. The greatest period of development was in the 1970s with more than one third of the total number of studies being published. During this period, and the previous decade, by far the most research was being conducted in the USA. This has not continued into the 1980s, however. Instead, many more people throughout Europe, Japan and Australia have started to examine the measure. The measurement of 'normal' voice and speech has remained prominent throughout the history of the LTS. However, it is interesting to note that the 1970s was the decade of research into individual speaker characteristics, particularly as applied to automatic speaker identification. This has fallen off considerably in the present decade, with more researchers turning to pathological voice and speech analysis, particularly throughout Europe, and to the measurement of long-term articulatory settings or voice qualities.

At present, we are able to classify the published works into 20 categories that are themselves grouped into three broad classifications. These are presented in Table 1 with an indication of the number of items in each category. Some, of course, fall into more than one category. As can be seen, researchers have seen the potential of LTS to provide quantitative descriptions in many aspects of voice and speech. We discuss below the relative utility of the LTS for a number of these applications. Before doing this, however, we focus on the LTS measurement itself.

GENERAL	
Methodology description or theory	77
Hardware implementations for measurement	12
Software implementations for measurement	1
Auditory scaling of frequency	8
DIRECT MEASUREMENTS	
Normal voice/speech analysis	69
Pathological voice/speech analysis	23
Analysis of speech of the deaf	1
Voice quality analysis	20
Analysis of the singing voice	6
Speaker characteristics	47
Language/dialect characteristics	37
Emotion/psychiatric characteristics	14
APPLICATIONS OF DIRECT MEASUREMENT	
Hearing aid designs	6
Telecommunications design	7
Architectural design	3
Automatic speaker verification	7
Automatic speaker identification	26
Clinical speech pathology	4
Clinical audiology	2
Clinical psychiatry	8

Table 1. Classification system

LTS MEASUREMENT

Computing the LTS has been an evolving process over the years. Early attempts required repeated scanning of the same segment of recorded speech with a succession of filters whose output was integrated in some way. As integration of the energy in each band was often accomplished using mechanical means, this stage of the analysis was both slow and error prone. It was in this kind of environment that the use of the 'chorus method' (Tarnóczy, 1956) was developed when multiple speakers were used in one recording session, and also when post-recording superposition of one part of the utterance on another part was used to reduce the length of the signal, and hence accommodate sufficient duration for statistical stability within the constraints of the measurement equipment. The development of banks of filters with electronic integrators enabled the LTS to be measured in one pass, but still using a fixed resolution. Until the advent of digital techniques to acquire, transform into the frequency domain, and manipulate results, most LTS measurement was dictated by available equipment rather than by the particular demand of the signal or the application.

The essential parameters of LTS measurement are 1) the time over which spectral energy is integrated, 2) the frequency range over which the energy is measured, 3) the frequency resolution of the measurement. Secondary issues relate to the way in which spectral energy is integrated, typically using either a linear or logarithmic energy value from sub-windows within the overall integration time, the number and type of speakers used, and the type of speech used (style of speech, method of collection, and whether any gating of the signal, such as separating voiced and voiceless sounds, is used). We will examine these issues in turn.

The integration time of the LTS computation is normally chosen to provide a stable estimate of the spectral shape that would be obtained over an unlimited time, that is when the statistical variation due to concatenation of a wide range of phonemes has stabilised. There are several alternate approaches used, not all of which will give comparable results. These range from extracting sounds in isolation and summing their spectra, to recording continuous speech over a sufficiently long period. Summing the spectra of isolated sounds is attractive because only a small amount of speech needs to be processed, but is also suspect as the proportion of each sound present in normal continuous speech is not accounted for, nor is the contribution of all the transitional components between the different component sounds. The former problem can be addressed by weighting the contribution of the individual sounds, but the latter problem cannot be rectified. An intermediate solution adopted in one study was to contrive word lists in which all the phonemes of the language were present in the correct proportion. This satisfies the first-order contribution of individual sounds and includes some, but only a subset, of the transitional components. Both of these 'economy' methods lack the contribution from the prosodic component of speech which influences spectral ripple due to intonation characteristics, and has some overall spectral shape effect through the use of various levels of stress. Even when using the straight-forward continuous speech accumulation method there are some differences. Several studies declare their integration time as the sum of all voiced segments, whereas others declare only the total elapsed time, within which hesitation and varying proportions of voiceless speech may occur. As it is clear that by judicious choice of material and speakers, reasonable estimates of the LTS can be achieved from relatively small portions of speech, it is not surprising to find that published averaging times vary from 10 seconds to 80 seconds. It should be noted, however, that very few studies give any indication of the stability of their LTS patterns. Where stability has been measured it has been found to vary with the passage of many months (Furui, 1974) and to be a factor that varies between speakers (Harmegnies & Landercy, 1988).

The frequency range and resolution are related variables. By reducing the range, fixed division of that range can be used to gain increased resolution. The most popular modern method is to use the fast Fourier transform (FFT) of sequential windows of the speech signal. Although it would be possible to use a single window utilising very large digital storage facilities, a window of 256 to 1024 points is commonly used giving 128 to 512 frequency values across the selected frequency range. Compared to early filter bank methods this allows very fine frequency resolution. Whereas analogue filter banks of constant Q, or their digital clones, naturally produce logarithmic frequency scales, the Fourier approach produces a linear scale.

Reviewing published studies, the finest frequency resolution reported is 5 Hz, and the coarsest, contiguous octave bands. In between, there are many variants. Uniform resolutions range from 5 Hz

to 250 Hz. Non-uniform resolutions normally follow either a fractional octave resolution, or an auditory scale such as critical band, bark or mel scale.

All of these scales can be derived from a raw fine uniform resolution scale produced by an FFT algorithm. Indeed, once these raw data are measured, they may be smoothed either to simulate a fixed-band filter bank of whatever coarser resolution, or by sweeping the frequency scale with a variable smoothing function to get a continuous fractional octave spectrum, or maybe a data-dependent resolution. An example of the latter would be to smooth at say twice the mean harmonic spacing of the voiced excitation, thus removing the excitation ripple from the LTS which is evident in monotonous voices.

It is clear that each of these variants will allow different global characteristics to be evident in the resulting spectrum. Fine linear resolution shows strong excitation ripple for monotonous voices. Excitation related smoothing will maximally emphasise peaks due to the dominant formant structure, third-octave smoothing will show formant related variation in the lower 2500 Hz, and one-octave smoothing will show only gross excitation differences at very low frequencies and otherwise emphasise a single major peak and a fairly uniform negative spectral slope above that peak.

The inclusion of voiceless speech for many voices has little effect. For others it can obscure some features in the 3000 Hz region and lift the over 4000 Hz region by a few dB.

In addition to LTS derived from various smoothings of Fourier analysis, or equivalent methods, the data in the LTS has been accessed by other methods which allow more efficient computation and reduced redundancy in the spectral description. Linear prediction analysis can be used to capture the shape of the spectrum within a given window to a variable degree of accuracy. As this analysis is based on an all-pole model of the speech spectrum it can be adjusted to reflect different degrees of detail in the spectrum in terms of the number of poles, or resonances, that are used in the model. This holds the potential for more data-dependent resolution while retaining economy of description. Partial auto-correlation analysis similarly gives an economical description of the spectrum.

Cross-correlation analysis starts with a Fourier analysis which is applied to a sequential set of windows throughout the speech. This set of time-sequential spectra is then transformed into a correlation matrix in which the degree of concurrent presence of spectral energy at different frequencies is encoded. This form of analysis gives more information than the simple LTS which ignores any characteristic combinations of spectral energy in temporal segments of the speech.

DERIVATIVE MEASUREMENTS

Once the spectrum has been produced, there are several types of measure that can be derived from the raw LTS. Being an amplitude spectrum measured over a selected frequency range, a number of major peaks are prominent. The nature of these peaks is highly dependent on the frequency resolution used (Millar, 1982). These peaks, and the overall spectral slope, or at least the slope from the first major peak, have been used by several studies, particularly early ones.

The proportion of energy lying in one frequency band to that in another has also been used to characterise certain pathological voice states. Thus, the proportion of energy in the band below 1 kHz to that above has been used to differentiate voices before and after voice therapy (Frøkjaer-Jenson & Prytz, 1976). This became known as the 'alpha-parameter'.

This type of manipulation is essentially a means of reducing the data. The raw LTS consists of numerous data points, particularly where a Fourier analysis of the sampled waveform is used. Often, a spectrum will consist of say 256 points equi-spaced across the frequency range, giving far more information than can be interpreted visually. Unless individual variation in the spectrum is the focus of the study, there is often a need to reduce the data. One is not restricted to reducing the spectrum to major characteristics such as peak values or slope, however. There are several multi-variate statistical techniques that handle large amounts of data at the one time and that also reduce them. Several studies have, as a result, manipulated the data using such statistics as factor analysis, hierarchical cluster analysis, and, in more recent times, three-mode principal components analysis. Such techniques allow all the information to be processed at the same time as it is reduced.

It can be readily appreciated that the application of this range of analysis techniques all under the umbrella of LTS makes valid interpretations of comparisons between studies which differ in technique, or which do not fully specify their techniques, hazardous. It is for this reason that a suitably classified and annotated bibliography of reported studies was deemed an important research contribution.

UTILITY OF THE LTS

When one considers the major application areas, no single computational technique stands out as most useful. Similarly, various frequency ranges and resolutions have been used by researchers in each area. Comparison, therefore, is indeed hazardous. We believed, however, that some indication of the utility of this measure should be given for the five major application areas.

Voice quality and pathological voice types are overlapping classifications, and are thus dealt with here together. There is a tendency for digital analysis to be used for these sets of studies, but this is simply because most have been conducted over the last decade. In general, laryngeal voice qualities, both pathological and normal, have been most readily differentiated by the LTS, although Esling (1983) has had some success with supralaryngeal settings. As a diagnostic tool for abnormal voice types, the LTS appears at this stage to have limited value. A number of researchers agree that its utility lies in its ability to document changes following treatment. Frequency ranges up to approximately 3 kHz seem to be most useful, while spectral detail can be reduced considerably and voice types still be differentiated.

The evidence for systematic language or dialect related differences in the early literature is slight. Even those studies employing a third-octave resolution, which might be expected to highlight formant variation due to different phonemic systems, do not produce significant results. More recently, however, several studies have reported distinctions that can be related back to both laryngeal and supralaryngeal voice quality differences linked to the phonemic system. Thus, a comparison of French and Dutch speakers (Harmegnies & Landercy, 1985) proposed that spectral differences found might be related to the presence of nasal vowels in the former. Similarly, the distinctions between Finnish and English spectra (Kiukaanniemi & Mattila, 1980) could be explained by the greater incidence of fricatives (and, therefore, pseudo-random noise similar to whisperiness) in English. Esling and Dickson (1985) related English dialect differences in the LTS to supralaryngeal voice qualities.

Most work on emotion or psychiatric states was conducted in the 1960s and 1970s using banks of bandpass filters with differing resolutions. In general, the range below 3 kHz has been shown to be the most useful, as have the finer frequency resolutions. Of the psychiatric states, severe depression is the one most studied. As with pathological voice types, the tendency has been to measure the voice before and after treatment, and once again, the LTS has been shown to be most useful in documenting change. Few studies have measured discrete emotions. Of those that have, the most clear distinctions are those between such emotions as anger and sorrow. These studies seem to indicate that the LTS reflects generalised affect dimensions such as arousal and pleasure, rather than discrete emotions as such.

As a measure of individual speaker identification the LTS has proved useful. This is one area that has tended to use large frequency ranges (up to 12500 Hz) and to analyse the correspondingly large amounts of data statistically. Discriminant analysis using a subset of the data as a reference followed by reclassification of the remainder has been used by several studies with marked success. In general, the finer the resolution, the better the result. Other derived measures such as a Euclidean distance metric or correlation matrices have also been used. A confounding variable in such studies is that the spectra of individuals may vary over time. Indeed, one study suggests that an average of several spectra from the same person is needed as a reference point. There have also been suggestions that a combination of measures such as cross-correlation and Euclidean distance enhance the possibility of speaker identification.

It can be seen from the foregoing analysis that LTS has proven useful in a number of application areas. Just how useful, however, is a little difficult to quantify given the hazards implicit in comparison across studies. When only a portion of the bibliographic information had been assembled it became apparent that, as with many areas of speech and language analysis, the confounding issues are both

multidimensional and multidisciplinary. In order to obtain an adequate grasp on these issues, a research tool was required to organise the data so that both its diversity and its unifying threads could be identified. An online bibliographic database was created which enabled us to integrate together all the relevant data while retaining the ability to divide and segregate application topics and analysis procedures. In the process, new structures within the published data have become apparent and have been incorporated into our classification scheme.

We therefore present this analysis and classification as a research tool (Pittam & Millar, 1988) and commend the process of establishing classified bibliographic databases in order to refine understanding and extend coverage of the literature beyond that normally accomplished by a literature review and reference list.

REFERENCES

Crandall, I.B. (1917) *The composition of speech* Phys. Rev. 10, 74-76.

Crandall, I.B., Mackenzie, D. (1922) *Analysis of energy distribution in speech* Bell System Tech. J. 1, 116-128.

Esling, J. (1983) *Quantitative analysis of acoustic correlates of supralaryngeal voice quality features in the long-time spectrum* In A. Cohen, M.P.R. v.d. Broecke (Eds.) *Abstracts of 10th International Congress of Phonetic Sciences, Utrecht*, pp. 363, (Foris: Dordrecht, Holland).

Esling, J., Dickson, B.C. (1985) *Acoustical procedures for articulatory setting analysis in accent* In H.J. Warkentyne (Ed.) *Papers from the Fifth International Conference on Methods in Dialectology* pp. 155-170, (Victoria: Dept. of Linguistics, University of Victoria).

Frøkjaer-Jenson, B., Prytz, S. (1976) *Registration of voice quality* Bruel and Kjaer Tech. Rev. 3, 3-17.

Furui, S. (1974) *An analysis of long-term variation of feature parameters of speech and its application to talker recognition* Electronics and Communications in Japan 57A, 34-52.

Harmegnies, B., Landercy, A. (1985) *Language features in the long term average spectrum* Revue de Phonetique 73-75, 69-80.

Kiukaanniemi, H.J., Mattila, P. (1980) *Long term speech spectra. A computerised method of measurement and a comparative study of Finnish and English data* Scandinavian Audiology 9, 67-72.

Millar, J.B. (1982) *Analysis of continuous speech for speaker characteristics* In J.E. Clark (Ed.) *Collected papers on normal aspects of speech and language*, pp. 225-252, (Speech and Language Research Centre, Macquarie University: Sydney).

Pittam, J., Millar, J.B. (1988) *Long-term spectrum of the acoustics of voice: An annotated and classified research bibliography* Indiana: Indiana University Linguistics Club.

Sivian, L. (1929) *Speech power and its measurements* Bell System Tech. J. 8, 646-661.

Stevens, S., Egan, J., Miller, G. (1947) *Methods of measuring speech spectra* J. Acoust. Soc. Am. 19, 771-780.

Tarnóczy, T. (1956) *Determination of the speech spectrum through measurements of superposed samples* J. Acoust. Soc. Am. 28, 1270-1275.