QUANTIFICATION OF SPEAKER VARIABILITY

J.B.Millar

Department of Engineering Physics
Australian National University

ABSTRACT - An approach to the quantification of the speaker dimension of speech is described. This utilises four phonetically motivated dimensions and a heirarchy of measurements in each, ranging from the macro-acoustics of long-term statistics to the micro-acoustics of the organisation of the syllable. A 33-speaker database of spoken Australian English is used to demonstrate the application of some of the macro-acoustic measurements and their importance for the development of robust speech processing for the Australian bionic ear project.

INTRODUCTION

Speech technology is built on foundations originating in many disciplines. Careful quantitative work in acoustic analysis, phonetics, pattern recognition, solid state physics, and electronic engineering has provided a body of knowledge and a range of techniques which allow speech technology devices to be implemented. The body of knowledge is however by no means complete. Much current speech technology navigates the gaps in our knowledge by means of statistical techniques. Variation in the acoustic signal for which we have no adequate model is treated as "noise", and systems which live with this noise perform better or worse depending on its level. One major component of such "noise" in a speech recognition system is the variation that is introduced by different speakers or by the same speaker speaking on different occasions. Ways of coping with "speaker noise" are occupying a significant proportion of current research and development in automatic speech recognition. An adequate model of speaker characteristics which is compatible with, complements, and interacts with linguistic models of speech is a fundamental component of a robust general purpose speech recognition system. This paper describes some initial work towards a more quantitative model of the speaker which may enable "speaker noise" to be managed by a partitioning of its range in a phonetically sensitive way, or more ambitiously by discovering greater structure within what we currently call noise.

We have selected four phonetically motivated dimensions which reflect the roles of diverse organs of the human body that are involved in the production of speech. Variance in speech timing, energy, excitation, and colour can be uniquely and independently individual. We now examine each dimension in turn to consider how some of the variables in these four dimensions may be measured in a robust manner, and then describe how such measurements subdivide the speakers in a 33-speaker database of spoken Australian English.

SPEECH TIMING

The timing of speech is influenced by many factors including the cognitive process of assembling the message, and the musculature of respiration and articulation. The way in which these factors influence speech timing can

be measured in different ways ranging from the macro-timing of overall duration of utterances, to the micro-timing of individual articulatory gestures. The overall duration of a lengthy utterance involving many respiratory cycles will encompass the summation of many different, and perceptually evident, timing strategies. Measurement of overall duration is technically trivial but provides a foundation on which to build the timing picture for a speaker. Beyond this foundation speech may be examined at the level of single respiratory cycles or breath groups. Durations of breath groups and their variation in a longer sequence of speech reveal temporal aspects of speech planning and breath resource management. These measures will also reveal the range of timing constraints that will propagate down to the production individual speech sounds. The lowest level of temporal units may be investigated by the measurement of inter-syllabic timing. Finer investigations are very difficult to perform using purely temporal measures such as "start" and "end", and boundaries are not always easy to distinguish. There is no plan in the current work to measure timing of units below the syllabic level.

SPEECH ENERGY

The acoustic energy inherent in speech sounds is controlled also by a number of organs. The movement of exhalatory muscles and the variation of constrictions at the larynx and other vocal tract sites influence the overall energy of a voice and the dynamic range of energy that contributes cues for the perception of stress and certain phonetic distinctions. The assessment of what constitutes significant energy to declare that speech is present, or of what constitutes significant change in energy to declare that a particular speech sound transition is in progress, underlies most other speech measurements. A baseline in this area is the distribution of energies present in a person's speech in a certain frequency band. There have been some isolated instances in the literature where the actual distribution of energy has been used to derive appropriate thresholds for subsequent analysis (Li et al., 1969; Blomberg and Elenius, 1970). Further, energy range characteristics have been shown to discriminate between a small sample of Australian and North American adult male voices (Millar and Wagner, 1983), and the energy distribution of speech, non-speech, vowels, nasals, and fricatives have been shown to be distinctive (Wagner, 1978).

In the current study a full-band energy histogram has been routinely produced for each one-minute speech passage analysed. As speech comprises periods of silence, inhalation of breath, frication, and phonation, the energies of all these components are included in the histogram. The major features for most speakers are peaks at energies roughly corresponding to silence and phonation. The ratio of these varies from speaker to speaker as does the additional contribution of inhalation, often indistinguishable from silence energy, and frication, which most often supplements the low energy skirt of the phonation energy distribution. For all practical purposes we have a two-peak histogram (figure 1). A first order interpretation of this relates the overall energy of the voice to the difference in energy between the two peak positions.

SPEECH EXCITATION

Speech excitation may be of several forms, a variety of kinds of phonation, frication caused by turbulence, or a mixture of frication and some kind of phonation. In the current analysis only phonation was measured by detecting excitation via transglottal electrical impedance. Instants of

glottal closure were determined by a simple algorithm acting on the impedance waveform. Histograms of time intervals between closures were produced. The first order model was to assume uni-modal symmetrical distributions, and to parameterise this model in terms of the mean and standard deviation of the distribution (figure 2).

SPEECH COLOUR

The acoustic wave at the termination of the vocal tract reflects the "colouring" of the excitation spectrum by the selective absorption of energy in the tract. Anatomical and habitual constraints on the shape of the vocal tract will give rise to long-term effects, whereas specific muscle gestures will characterise the articulation of individual sounds.

Speech colour has been measured at two levels. Firstly, we measured the long-term spectrum of one-minute passages of speech and viewed the results at three levels of resolution (Millar, 1982). Secondly, we modelled the resonance patterns in syllables in order to express, in a speaker-specific way, the expected resonance trajectories between specific consonants and vowels (Clermont and Millar, 1986).

THE DATABASE

Our database of spoken Australian English (O'Kane, Millar & Bryant, 1982) comprises 15 male and 18 female speakers whose accent may be broadly classified as General Australian. In this paper we focus on the analysis of three reading passages designed to last one minute, which were recorded over a period of approximately 6 weeks from all 33 speakers. The recordings from each speaker were separated by at least 2 weeks in most cases. Passage A is an expository discourse from a popular scientific text (with some scientific terms omitted), passage B is a narrative from a childrens' book including a fair amount of dialogue, and passage C comprises two short passages often used in speech research - the fable of the "North wind and the Sun" [C1], and the "Rainbow" passage [C2].

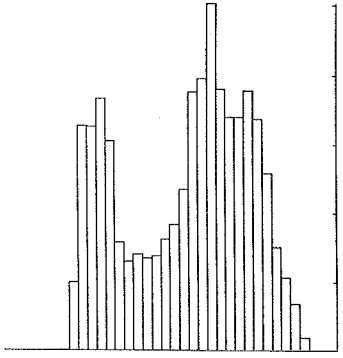

Figure 1. Distribution of occurences of energy levels in a One-minute passage.

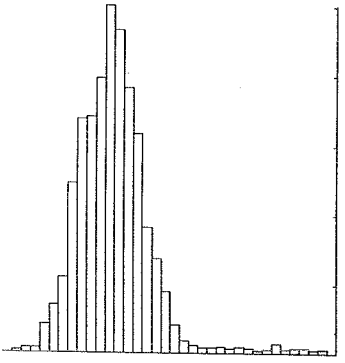

Figure 2. Distribution of intervals between glottal closures in a One-minute passage.

POPULATION CHARACTERISTICS

Before looking at individual speaker variation across the reading passages, we gauge the overall population variance in timing and excitation.

The overall durational analysis showed strong consistency over all passages. The standard deviation of durations across all speakers was approximately 10% in nearly all cases. This result holds separately for the male and female sub-populations, and there was no significant difference between the overall durational characteristics of male and female sub-populations. The raw data are given in table 1.

Table 1. Overall Durations of reading passages.

Passage	Male mean	St.Dev.	Female mean	St.Dev.
A	66.24 sec	9.1%	69.44 sec	10.3%
B	69.57 sec	9.6%	70.78 sec	12.5%
C1	32.74 sec	9.5%	33.28 sec	8.6%
C2	31.19 sec	10.4%	31.85 sec	10.3%

The excitation analysis was performed on all those speakers for whom impedance waveforms could be measured. The population-wide results for the residual male and female subpopulations are given in table 2.

Table 2. Overall excitation characteristics in terms of parameters of the distribution of intervals between glottal closures.

Pass-	MALE				FEMALE			
age.	Median	Mean	St.Dev.	Spread	Median	Mean	St.Dev.	Spread
A	9.5ms	9.6ms	1.64ms	1.6ms	5.0ms	5.1ms	0.36ms	0.9ms
B	8.9ms	9.0ms	1.27ms	1.8ms	5.1ms	5.2ms	0.37ms	0.9ms
C1	9.4ms	9.5ms		1.6ms	5.2ms	5.2ms		0.8ms
C2	9.4ms	9.5ms		1.7ms	5.2ms	5.2ms		0.8ms

CLASSIFICATION OF SPEAKERS VIA A FEW MACRO-ACOUSTIC CLASSIFIERS

Several macro-acoustic measures have been applied in order to classify the speakers in a multi-dimensional space involving two of the dimensions mentioned above. Macro-timing analysis revealed several revealed several different types of speakers - those who are consistently slow, medium, or fast speakers, and those who vary their speed in a way that is influenced by the material. The overall duration of passages A, B, C1, and C2 were measured and each speaker classified with respect to the population mean for each passage. Those within 0.5 standard deviations of the mean were classified as medium rate speakers, those more than 1.2 standard deviations from the mean were classfied as fast or slow speakers, and the two remaining groups were classified as medium-fast or medium-slow speakers.

Macro-excitation analysis has revealed speakers who exhibit diverse combinations of mean level of voice pitch and range of intonation. Within our sample the females tend to have a proportional relationship between pitch and intonation range (figure 3a). The males however, show a more complex relationship with the most quantitatively monotonous voices being

in the middle of the mean pitch range (figure 3b).

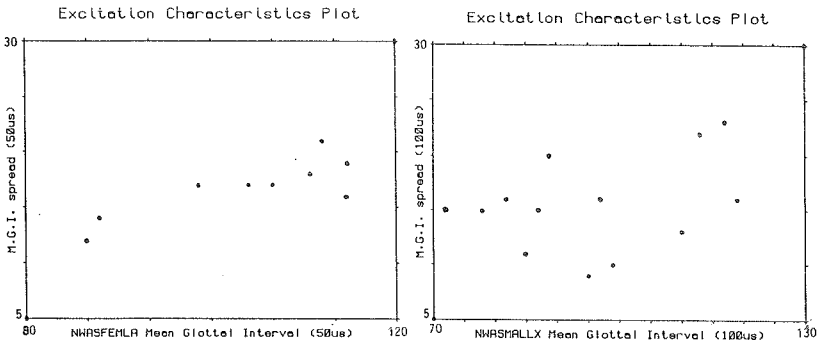

Figure 3. Scatter plots of mean glottal interval vs spread of glottal intervals in a half-minute passage for (a) females, and (b) males.

APPLICATION OF CLASSIFICATIONS TO SPEECH PROCESSING FOR THE BIONIC EAR

This preliminary level quantitative analysis has recently been applied in the development of new speech processors for the Australian bionic ear project at the University of Melbourne. Throughout the world there is considerable variance in the kind of speech processing used for translating the acoustic speech signal into a form that may be used to directly stimulate the auditory nerve (Millar et al, 1984). The current Australian device uses a feature extraction method, presenting fundamental frequency and spectral resonance information. The techniques to extract and transfer this information are being continually updated in terms of philosophy and implementation. Typically, a new approach is tested against an old approach to judge their differential results. One of several variables that can jeopardise the efficiency of this process is that of individual speaker characteristics. The use of "live-voice", while enhancing patient confidence and interest, brings its problems of repeatability and maybe bias, whereas the use of a single recorded voice brings its problems of lack of variance that is normal in the speech of one individual or across a community of speakers. It is clear that most users of the bionic ear speech processor will wish to use their device with a variety of speakers. If development of feature extraction techniques is to produce algorithms which will behave in a predictable way across a wide cross-section of the population of speakers, then that development must be exposed to this wide cross-section. The speech literature is littered with short-lived ideas that worked well with their "master's voice". It is therefore necessary to thoroughly test algorithms for such speech processing on representative data.

Four speakers have been extracted from the database who conform to the following criteria:

1. They continue to be available for further samples.

2. As a group they represent both male and female speakers having mean fundamental frequencies approximately one standard deviation above and below the male and female population averages.

3. They have a speaking rate that is below the population average.

4. They have the highest range of fundamental frequencies consistent with the above requirements.

These criteria provide speech samples which may be extended [1], which may be used directly or in segmented form for perceptual experiments [3], which represents a representative range of fundamental frequencies [2], and in which there is plenty of speech dynamics [4]. It was felt that these basic criteria provided, within the scope of four voices, a reasonable spread of speaker variance typical of the normal population. Other factors such as phonation type and vocal tract length may well be applied as analysis of the database proceeds.

CONCLUSIONS

The use of a structured and phonetically motivated set of measurements for quantifying speakers has been suggested. The scheme has been applied in part to a 33-speaker database of spoken Australian English, and preliminary results have been outlined. It has also been shown how such a technique can provide a small number of representative speakers for the evaluation of new developments in speech technology.

REFERENCES

BLOMBERG,M., ELENIUS,K. (1970) "Statistical analysis of speech signals", QPSR/4 Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, 1-8.

CLERMONT,F., MILLAR,J.B. (1986) "Multi-speaker validation of coarticulation models of syllabic nuclei", Proc. ICASSP-86, 2671-2674.

LI,K.-P., HUGHES,G.W., HOUSE,A.S. (1969) "Correlation characteristics and dimensionality of speech spectra", J.Acoust.Soc.Amer. 46, 1019-1025.

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", Speech & Language Research Centre, Occasional Papers, Macquarie University.

MILLAR,J.B., TONG,Y.C., CLARK,G.M. (1984) "Speech processing for cochlear implant prostheses", J.Speech.Hear.Res. 27, 280-296.

MILLAR,J.B., WAGNER,M. (1983) "The Automatic Analysis of acoustic variance in speech", Language and Speech, 26, 145-158.

O'KANE,M., MILLAR,J.B., BRYANT,P. (1982) "A database of spoken Australian English: Design and Collection", Technical Note No.6, School of Information Sciences, Canberra College of Advanced Education.

WAGNER.M. (1978) "The application of a learning technique for the identification of speaker characteristics in continuous speech", Unpublished Ph.D. Thesis, Australian National University.