QUANTIFICATION OF A MULTI-SPEAKER DATABASE OF SPOKEN AUSTRALIAN ENGLISH

J.B.Millar

Department of Engineering Physics
Australian National University
Canberra, Australia, 2601

ABSTRACT

Measurements for the quantification of the speaker dimension of a database of spoken language are presented. These measurements are structured in two dimensions - utterance extent and acoustic feature. Utterance extent ranges from the macro-acoustics of long-term statistics to the micro-acoustics of the organisation of the syllable, while the acoustic features cover the phonetically motivated areas of Energy, Timing, Excitation, and spectral Colour. An example is given of the application of a small sub-set of these measurements to extract representative speakers from a 33-speaker database of spoken Australian English for the development of robust speech processing for a cochlear implant project.

INTRODUCTION

The acoustic-phonetic description of the speaker dimension of speech has received some solid attention in recent years [2,5]. These studies have approached the issue from the standpoint of phonetics with acoustics used to give quantitative backup to the phonetic judgements. In this paper, the starting point is acoustics, and the aim is to present a hierarchy of measurements that can specify speaker characteristics with increasing refinement, and to apply a subset of these measurements to a specific practical task.

Much work in acoustical speaker characterisation has in view high resolution speaker discrimination for a variety of purposes. One route to this goal is the direct route of looking for discriminating cues. This route however begs the question of the density of the speaker space from which test speakers are selected. My route is to first of all map out the speaker space and, with progressive refinement, plot out the trajectory of speakers within that space as they use language in different ways and over periods of time. This approach is motivated by the belief that speech technology needs an adequate model of speaker characteristics which is compatible with, complements, and interacts with linguistic models of speech.

I 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 spectral colour can be uniquely and independently individual. Each of these dimensions may be examined at more than one level of utterance extent (Figure.1). As speech is implemented by sets of short-term articulatory gestures superimposed on longer term breath resources and even longer term patterns of articualtory status, my analysis strategy aims to see the short-term against the background of the long-term, and to use the long-term to constrain the short-term descriptions. Analysis across all the four feature domains and the timescales is in progress but only a subset are presented by results of its application to the speakers in a 33-speaker database of spoken Australian English.

	LONG TERM	BREATH GROUP	SYLLABLE
ENERGY	Long-term Energy	Multisyllable Pattern	Contour
TIMING	Overall Duration, and Breath Group Structure	Inter-syllabic Intervals	Duration
EXCITATION	Long-term Excitation Frequency Distribution	Intonation Pattern	Contour
COLOUR	Long-term Spectrum		Contour

Figure 1. Analysis structure for Quantification of of Speakers (see text).

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 timing strategies. Measurement of overall duration is technically trivial but provides a foundation on which to build the timing picture for a speaker. The direct contribution of pauses for inspiration (respiratory) may be removed by the decomposition of longer utterances into single respiratory cycles or breath groups. Durations of breath groups and their variation reveal temporal aspects of speech planning related to the semantics and syntax of the utterance, and to the management of breath resources. This individually determined mapping of linguistic content onto the time course of a breath carrier will to some degree determine the repertoire of prosodic expression which may be used - each with its own breath resource penalty in addition to prosodic pattern and pattern continuity constraints. The complete picture of prosody management is undeniably complex but is here identified as a speaker variable begging quantification. At the simplest level syllabic rate and intersyllabic interval within breath groups provide a more accurate timing description than overall syllabic rate. Accurate temporal identification of the syllable provides the foundation for syllabic modelling. Sub-syllabic temporal segments are not attempted as this area is best catered for by coarticulatory models of the syllable.

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 many speech measurements. Energy range characteristics have been shown to discriminate between a small sample of Australian and North American adult male voices [4], and the energy distribution of speech, non-speech, vowels, nasals, and fricatives have been shown to be distinctive [7].

In the current study a full-band long-term 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 2). An initial interpretation of this histogram relates the overall energy of the voice to the difference between the two peak positions and may be considered a speaker characteristic if all recordings were made in the same environment with a constant ambient noise. Further speaker characteristics may be attributed to the shape and relative size of the energy peaks but have proved to be difficult to extract with any reliability.

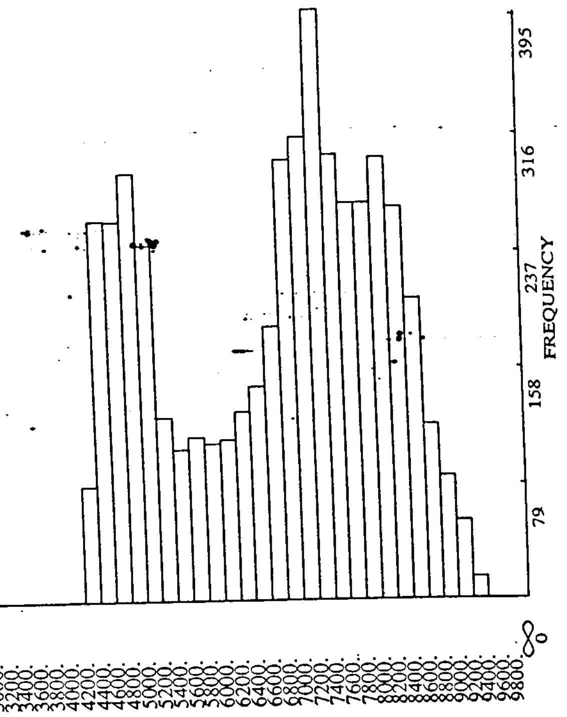

Figure 2. Distribution of occurences of energy levels in One-minute passage.

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 3).

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 [3]. 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 [1].

THE DATABASE

Our database of spoken Australian English [6] 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, each 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 scient~~ific 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).

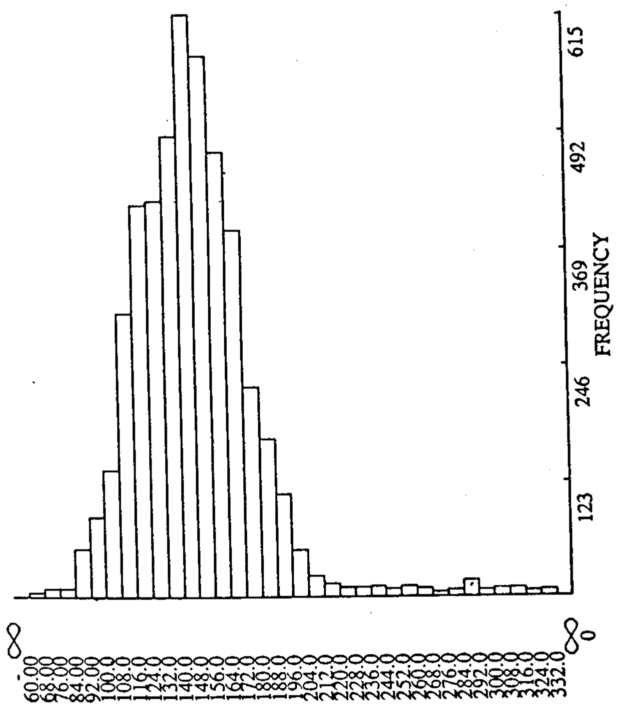

Figure 3. Distribution of intervals between glottal closures in a One-minute passage

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.

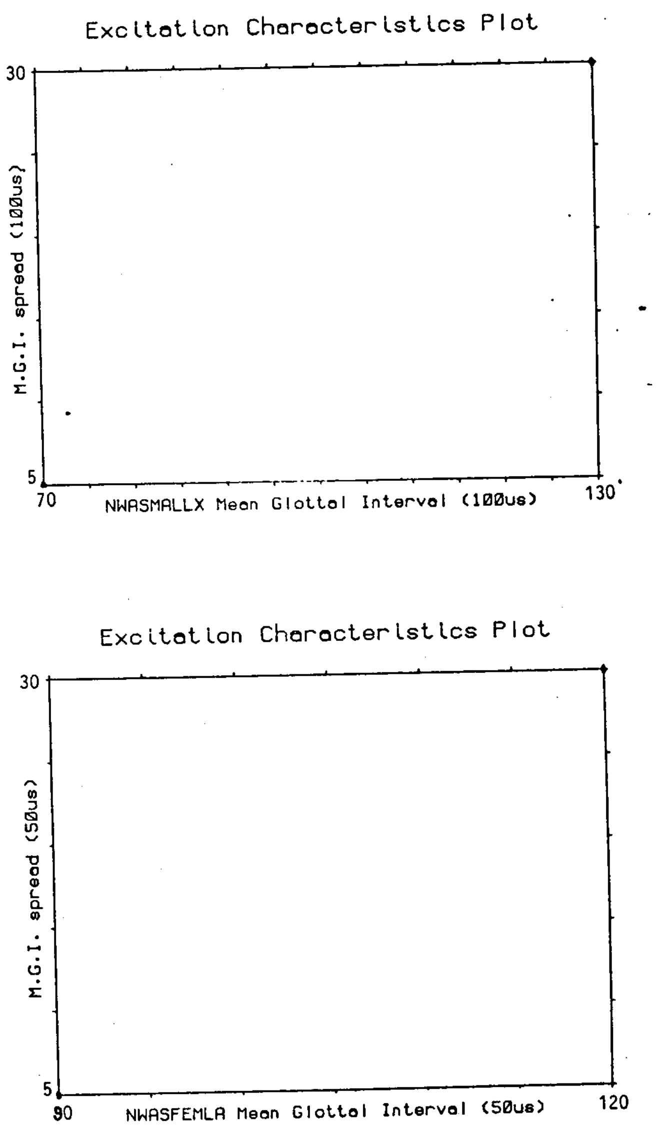

Figure 4. Scatter plots of mean glottal interval vs spread of glottal intervals in a 30s passage for (a) males, and (b) females.

Macro-timing analysis revealed several different types of speakers - those who are consistently slow (>1.2sd above mean), medium (within 0.5sd of mean), or fast (>1.2sd below mean) speakers, and those who vary their speed in a way that is influenced by the material.

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. This 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 4a). The males however, show a more complex relationship with the most quantitatively monotonous voices being in the middle of the mean pitch range (Figure 4b). The mean standard deviation of inter-glottal closure intervals for males and females over all passages is in the range 15-20%. This is equivalent to a total range of less than 1 octave.

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

	MALE				FEMALE			
	Med- -ian (ms)	Mean (ms)	SDof Mean (ms)	Mean ofSD (ms)	Med- -ian (ms)	Mean (ms)	SDof Mean (ms)	Mean ofSD (ms)
A	9.5	9.6	1.64	1.6	5.0	5.1	0.36	0.9
B	8.9	9.0	1.27	1.8	5.1	5.2	0.37	0.9
C1	9.4	9.5		1.6	5.2	5.2		0.8
C2	9.4	9.5		1.7	5.2	5.2		0.8

CLASSIFICATION OF SPEAKERS

Macro-acoustic analyses in the domains of timing and excitation have been used to select four speakers whose speech can be used for testing speech processing algorithms for the Australian Cochlear Implant project at Melbourne University. It is necessary to thoroughly test algorithms for such speech processing on natural data that is representative of the kind of speech that the implantee is likely to receive.

The selected speakers conform to the following criteria:

1. 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.

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

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

These criteria provide speech samples which may be used directly or in segmented form for perceptual experiments (2), which represents a representative range of fundamental frequencies (1), and in which there is plenty of speech dynamics (3). 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 that is relevant to this specific application. Other factors such as phonation type and vocal tract length may well be applied as analysis of the database proceeds.

SUMMARY AND DISCUSSION

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.

Such a hierarchy of measurements for progressively refined quantification of the speaker-space are most important for the development of speech technology which admits the use of multiple speakers. Speaker-independent automatic speech recognition needs to be tested against speaker groups which are evenly spread, or which have defined clustering characteristics. Only in this way will comparisons between techniques be valid. Conversely, many distinct voices in multi-voice speech synthesis may be produced when distributed evenly over the perceptual equivalent of the analytic speaker-space.

REFERENCES

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

[2] LAVER, J.D.M. (1980) "The Phonetic Description of Voice Quality", Cambridge University Press: Cambridge.

[3] 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.

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

[5] NOLAN, F.J. (198) "The Phonetic Bases of Speaker Recognition", Cambridge University Press: Cambridge.

[6] 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.

[7] 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.