STABILITY OF LONG TERM ACOUSTIC FEATURES

J.Bruce Millar

Computer Sciences Laboratory
Research School of Physical Sciences
Australian National University

ABSTRACT - This paper is a progress report on an ongoing study of speaker characteristics in a number of acoustic feature domains and a number of temporal domains. Variations in the long-term analysis of timing, energy distribution, fundamental frequency distribution over a three month period for 33 speakers of Australian English are presented. These data are based on 5 reading passages of a nominal duration of one minute.

INTRODUCTION

An understanding of the long-term statistical nature of the energies and frequencies present in the speech signal assumed increasing importance as telecommunications became established in the early part of this century. Only with this knowledge could telecommunication channels be designed efficiently. The development of new coding schemes based on this basic knowledge continues today as new technology makes the exploitation of this knowledge economically feasible.

More recently the attention of speech researchers has been focussed on the processing of the speech signal by a machine rather than by a human as in most telecommunications applications. As machine-reading of speech is aimed primarily at linguistic information, its shorter-term segmental characteristics of phones, syllables, words, phrases, and sentences have tended to dominate the investigation of its acoustic form. It is apparent however that one of the major stumbling blocks to gaining an adequate taxonomy of the acoustic form of these linguistic entities is the degree of variability that exists between speech sounds spoken in different contexts, in different speaking styles, and by different speakers. A distinct hole exists in our understanding of the way in which these sources of variability may be integrated into a taxonomy of speech acoustics other than in a very general statistical fashion. One approach to increase our understanding of speaker characteristics is to apply techniques similar to those used to characterise speech in general to the speech of individual speakers.

The characterisation of individual speakers, however, needs to occur within a multi-faceted model of speech performance which acknowledges the contribution of the diverse articulatory apparati of the speech producers, the scope of auditory normalisation of phonetically salient information from diverse acoustic signals in the speech perceivers, and the interaction between the nature of the speech task in progress with the structure of the language at lexical, syntactic, and semantic levels. All of these components need to be tied together in a cohesive framework. The present study is a contribution to foundations for a speaker-sensitive acoustic framework for Australian English.

I am conducting a carefully controlled analysis to establish foundational data on spoken Australian English and robust algorithms to derive these data. Foundational data are important because what is currently known is rather fragmentary and not an adequate base on which to build phonologically sensitive automatic speech recognition (ASR) technologies. Robust algorithms are clearly necessary to establish the internal integrity and extendability of a quantitative description of the acoustics. It is to be expected that in certain dimensions our national speech will differ little from that of related English speaking dialect groups. In other dimensions such as the diphthongisation of vocalic nuclei and the relatively unexplored area of styles of prosody, quite significant differences may be expected. It is important for the rigour of our own understanding to build our increasing knowledge into an adequately extensive framework which will then contain these local characteristics in their global context.

When performing speech analysis in preparation for ASR many seemingly ad hoc settings of filters, thresholds, time periods, and other facets of the analysis procedure, have to be made. These 'pa-

	LONG TERM domain	BREATH GROUP domain	INDIVIDUAL SYLLABLE domain
ENERGY features	Long-Term Energy	Multi-syllable pattern	Contour
TIMING features	Overall Duration, and Breath Group Structure	Inter-syllabic intervals	Duration
EXCITATION features	Long-Term Excitation Frequency Distribution	Intonation Pattern	Contour
COLOUR features	Long-Term Spectrum		Contour

Figure 1: Acoustic features and time domains of analysis

rameters' of the analysis process are chosen to generate a desired result in terms of classification of some aspect of the speech signal, and the reason for a 'parametrised' approach rather than using fixed values is that speakers do not all perform in similar ways: similar, that is, either to other speakers or to previous utterances of their own. It is therefore important to know which parameters of a speaker can be set and left as stable, and which need to be more constantly monitored and updated in some way, both within the speech of a single speaker and across the speech of different speakers.

I have therefore been concerned to explore some of the most basic acoustic features of speech on a number of time-bases (Millar, 1982; 1986; 1987). This paper will focus on currently completed analysis in the long-term domain (figure 1), and will extend earlier presentations by examining the variation that occurs over time and speech materials.

LITERATURE

The degree of variance that exists in the speech of individual speakers is a subject that has received scant attention in the literature, but there have been some notable exceptions to this trend. Furui et al.(1972) examined the temporal variance in the long-term spectrum within the context of a talker recognition system. He showed that this long-term measure remained stable for different speakers for periods of 2-3 days up to 3 weeks, but beyond that longer-term variation was found. Nolan (1983) emphasised that the task of characterising a speaker in terms of their speech performance was one of identifying the bounds of their habitual speech behaviour. This may be regarded as identifying their susceptibility to variance in speech performance in a range of circumstances. Harmegnies and Landercy (1988) recently coined the term 'coherence' to describe the consistency with which a speaker performs. They showed that 'coherence' is a strong speaker variable in the domain of the long-term spectrum. The related concepts of the 'coherence' and 'susceptibility to variance' of a speaker's acoustic performance are central to the philosophy of the present study. This paper focusses on susceptibility to variance in timing, in voiced energy, and in excitation frequency.

DATABASE

Our database of spoken Australian English (O'Kane et al., 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 five reading passages of nominal duration one minute each. The readings took place over a period of approximately 3 months with no less than one week between readings. All 33 speakers completed the readings with just one speaker missing out a few words (3%) in one passage. Passage A is an expository discourse from a popular scientific text (with some scientific terms omitted), passage B is a procedural discourse aimed at children, passage D is a narrative discourse containing approximately 38% dialogue, passage H is a narrative discourse containing a very small amount (about 1%) of dialogue, and passage J is a narrative discourse containing no dialogue. The passages were recorded by most speakers in the order given above.

ANALYSIS

The speech data were collected on two channels, the pressure waveform (PW) from a flat frequency response microphone, and the larynx impedance signal (LX) from a laryngograph. Both signals were sampled at 20000 samples per second. The LX signal was used wherever possible to detect the voice excitation frequency and to signify whether the speech in a particular 3.2ms frame was voiced or unvoiced. As reliable measurement of the LX signal was not possible for all speakers reading all passages, an alternative procedure, applying a cepstrally based 'pitch and voicing detector' (ILS/API) to a downsampled version of the PW signal, was used when the LX signal failed. This failure, due mainly to excessive noise in the signal, occurred in 4/75 male speaker passages, and in 25/90 female speaker passages.

Following the detection and measurement of the voiced excitation characteristics, the speech was subjected to Fourier analysis. This provided the basis for the measurement of frame-by-frame energy and an estimate of frame-by-frame loudness. The former was also used to compute the long-term energy distribution within each passage and the latter was used to perform syllabic and breath group detection (see figure 1). Apart from breath group analysis, the breath group detection was also used in the long-term analysis to separate non-speech and silent intervals from continuous speech sequences. The speech signal within breath groups was then described as 'speech' and that outside of breath groups as 'background'. Within the 'speech' data the further distinction of 'voiced' and 'unvoiced' is made on the basis of LX or ILS/API analysis mentioned above.

These analyses enable us to examine statistics of several long-term characteristics of the speaker in the area of timing, voiced energy, and excitation frequency. The Fourier magnitude spectrum is also used to compute the long-term spectrum but currently no data reduction has been performed on this.

RESULTS

We report the results of these analyses in four ways. First, we specify the overall distribution of the values of a specific acoustic feature for our sample of the two sex-differentiated groups - the F-population and the M-population. These values 'locate' our sample of the speaker population in terms of other published values - although there is no place in this paper to make these comparisons explicit. Second, we examine the degree of variation that exists within and between the M and F populations - that is whether speakers and/or sexes are differentiated by their performance along that particular acoustic feature domain. Third, we examine whether the distributions have any significant variance across time between recording sessions or with respect to the one different passage (D) which contained a substantial amount of dialogue (DIAL). Fourth, we examine the individual speaker standard deviations (ISSD) across the passages.

TIMING

The overall duration (ODR) is the actual time taken to read the passage including minor hestitation or respiration pauses. Gross hestitation, repetition due to error correction, asides, coughs, etc were removed. The background duration (BDR) is the proportion of ODR that is spent in non-speech activity, including hesitation and respiration. The voiced percentage of speech (VPS) is the percentage of continuous speech (ODR minus BDR) that is voiced.

As no passage was repeated we cannot report on the stability of ODR of individual speakers over time, however we can report on the population variance in ODR across the different passages. There was no significant difference between male (stdv=7.8-10.1% of mean) and female (stdv=7.8-12.5% of mean) ODR variance across the passages. Likewise there was no significant difference between the ODR means of the M and F populations on a per-passage basis even though the F-population consistently produced longer durations than the M-population in every situation, producing a significant M-F difference over all the data (two-tailed t, at 2% - Table 1). There is therefore mild evidence to characterise our F-population as 'slower' speakers than the M-population.

The BDR analysis also revealed a mildly significant difference (2%) between the M and F populations over all passages but no clear significance on a per-passage basis. The overall mean BDR was 22.3% for males, and 19.8% for females. Thus while the ODRs of the F-population are consistently greater than those of the M-population, their BDRs are consistently lower. Thus on average the females spent longer in speech articulation than the males faced with the same speech tasks. The spread of BDR means was larger for the F-population, but also their intra-speaker variation (BDR ISSD mean) was significantly larger. One other area of significance noted in BDR measurement was the impact of the dialogue component. There was a significant difference in BDR between passage D and all other passages for males and for all but one passages for females.

TABLE 1.

TIMING	MALE		FEMALE		SIGNIFICANCE
	mean	stdv	mean	stdv	(two-tailed t-test)
ODR mean	70.28s	7.10s	73.30s	8.12s	t=2.503 [2%]
BDR mean	22.30%	5.65%	19.76%	7.28%	t=2.450 [2%]
ISSD	4.27%	1.66%	5.90%	2.12%	t=2.348 [5%]
DIAL	(4 passages diff)		(3 passages diff)		
VPS mean	63.92%	5.46%	55.83%	8.00%	t=7.388 [0.1%]
ISSD	3.84%	1.53%	5.62%	1.79%	t=2.943 [1%]
DIAL	(positive correl)		(positive correl)		with % dialogue [NS]

The VPS measure revealed a highly significant 14% difference between the M and F populations. The greater duration of voiced speech in the males was significant also on a per-passage basis. The strength of this effect immediately suggested an analysis artefact owing perhaps to the failure of the voicing detectors to work properly with higher pitched voices. To date such an artefact has not been found. In particular the low VPS for females is not caused by the greater use of ILS/API for females. A thorough review of the performance of voicing detection from the LX signal for female voices is currently in progress. Comment on the significantly different VPS ISSD values must be deferred until this review has been completed.

VOICED ENERGY

The mean value of voiced speech energy (VEM) is expressed as the ratio (dB) of the mean of the voiced energy distribution to the mode of the 'background' energy distribution. It is therefore disassociated from any gain variation between speakers and recording sessions as all recordings where made in the same acoustic environment. The spread of voiced energy (VES) is one standard deviation of the voiced energy distribution.

There was no significant VEM difference between the M and F populations (Table 2). The overall mean was approximately 22 dB within an expected range of 17-27 dB. The male and female results for ISSD of VEM had very similar mean values (1.8dB), but the females showed more variation: ISSD values from 0.8dB to 3.5dB were measured. Therefore the VEM values for certain speakers could be predicted 95% of the time to an accuracy of 1.6dB, while for others the uncertainty could be as much as 7dB. For the former thresholding on energy could be fixed, but for the latter constant empirical checks would need to be made. Population wide VEM values did not vary significantly across time or reading passage style.

TABLE 2.

ENERGY in dB.	MALE		FEMALE		SIGNIFICANCE
	mean	stdv	mean	stdv	(two-tailed t-test)
VEM mean	21.68	2.40	22.18	2.40	t=1.324 [NS]
ISSD	1.83	0.57	1.84	0.74	t=0.041 [NS]
DIAL	Not significant		Not significant		
VES mean	7.76	0.84	6.77	1.09	t=6.394 [0.1%]
ISSD	0.63	0.34	0.86	0.26	t=2.132 [5%]
DIAL	3 passages diff		4 passages diff		

A very significant VES difference was found between the M and F populations, with male speakers having an average value which is 1dB larger than that for female speakers per standard deviation of spread. This means that our male speakers on average use an extra 4dB in voiced energy dynamic range above that used on average by our female speakers. Individual variation in VES is 0.63dB [M] and 0.86dB [F]. These too are significantly different indicating the greater stability in personal style of use of energy range in the males than in the females. This observation is reinforced by an analysis of the difference in VES between passage D and the other passages. Both male and female showed significant differences but the female differences were more highly significant.

EXCITATION FREQUENCY

In each passage the excitation frequency was tracked throughout the voiced speech and expressed in terms of a mean value (EFM) and a one standard deviation spread value (EFS). These values were computed on a 16th-tone scale with an origin at 55Hz. The EFM values were converted back into Hertz but the EFS left as tonal values according to the convention in the literature.

TABLE 3.

EXCITATION	MALE		FEMALE		SIGNIFICANCE
	mean	stdv	mean	stdv	(two-tailed t-test)
EFM mean	104 Hz	1.3 t	190 Hz	0.7 t	exclusive ranges
[Hertz range]	[77-141 Hz]		[161-225 Hz]		exclusive ranges
ISSD	.33 t	.16 t	.27 t	.13 t	t=1.248 [NS]
DIAL	Not Significant		Not significant		
EFS mean	1.4 t	.37 t	1.2 t	.33 t	t=3.546 [0.1%]
ISSD	.22 t	.13 t	.22 t	.10 t	t=0.137 [NS]
DIAL	3 passages diff		3 passages diff		

The male and female population means for EFM are stable over the three month period, and even though the male mean EFM value for passage D was raised, it was not significantly different. The average individual male variation in EFM over time had a standard deviation of one third of a tone, and this average variation itself had a standard deviation of a sixth of a tone (Table 3). The maximum likely individual variation is therefore about two-thirds of a tone, and the minimum close to zero. The actual distribution of variations was somewhat skewed with a maximum close to two-thirds of a tone and a minimum of one-eighth of a tone. The equivalent female ISSD values had a very similar range but a mean level of variation close to a quarter of a tone.

The mean shift in the male EFM mean for passage D with respect to the individual speaker means was 0.85 individual speaker standard deviations, but with a standard deviation of 0.75 ISSD. Out of the 15 male speakers, two speakers' EFMs shifted by approximately 1.7 ISSD, seven by 1.0-1.5 ISSD, and the remaining six had little reaction.

Examination of mean EFS values across the M and F populations revealed a highly significant differences. The expected range of EFS means was 0.7-2.1 tones for males and 0.5-1.9 tones for females. As the EFS itself is a standard deviation, these results translate into a range of 'pitch dynamics' from the most monotonous female voice having 2 tones (one third octave) variation, to the most tonally dynamic male voice having 8.4 tones (close to one and a half octaves) variation. Although the variability over speakers and sexes is considerable and population wide EFS mean is significantly increased for passage D, individual speaker variation over the passages was less marked. Both males and females varied in EFS with individual standard deviations ranging from near zero to a semitone, with a mean of .22 tones.

CONCLUSIONS

The overall duration of the readings varied across speakers with a standard deviation of approximately 10%. Differences in timing between males and females were mildly significant but may be attributed to the sampling of the wider population. The amount of time spent in speech articulation is clearly sensitive to the degree of dialogue in the passage. Individual differences in speech articulation time over time and passage style show an average standard deviation of about 5% of total duration around a mean value of 80% of the total duration.

Mean voiced energy measurements indicate considerable inter-speaker variability in the level and consistency of voiced energy, but not on the basis of sex. Fixed thresholds on energy are justified for some speakers but not for others. Differences in the spread of voiced energy indicate that thresholding by using a fixed value below an empirically determined maximum energy for a given speaker could generate an analysis artefact between the processing of male and female speech, even apart from individual variations.

Excitation frequency measurements indicate the expected largely separate distributions for mean value for male and female, with ranges of about one octave, and about half an octave, respectively. The individual range of frequency showed a 4:1 ratio between the most monotonous to the most dynamic. The importance of these values relates to the possibility of speaker specific speech information coding strategies as they effect the spectral sampling of the vocal tract (Millar and Wagner, 1983).

REFERENCES

Furui,S., Itakura,F., Saito,S. (1972) *Talker recognition by long time averaged speech spectrum,* Electronics and Communications in Japan, Vol.55A, No.10, 54-61.

Harmegnies,B., Landercy,A. (1988) *Intra-speaker variability of the long-term spectrum,* Speech Communication, Vol.7, 81-86.

Millar,J.B. (1982) *Analysis of Continuous speech for speaker characteristics,* In "Collected papers on normal aspects of speech and language", J.E.Clark (ed), 225-252.

Millar,J.B. (1986) *Quantification of speaker variability,* Proc. 1st Australian Conf. Speech Science and Technology, Canberra, 228-233.

Millar,J.B. (1987) *Quantification of a multi-speaker database of spoken Australian English,* Proc. XIth Int. Cong. Phonetic Sciences, Tallinn, 245-248.

Millar,J.B., Wagner,M. (1983) *The automatic analysis of acoustic variance in speech,* Language and Speech, Vol.26, 145-158.

Nolan,F. (1983) *The phonetic bases of speaker recognition,* Cambridge University Press: Cambridge.

O'Kane,M., Millar,J.B., Bryant,P. (1982) *A database of spoken Australian English: Design and Collection,* Technical Note No.6, Canberra College of Advanced Education.