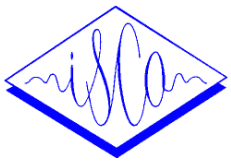

ESCA-NATO Workshop on
Speech under Stress
Lisbon, Portugal
September 14-15, 1995

Effects of Anxiety in Visual and Audio Speech Databases

J. Thompson & J. S. Mason

Speech Research Group, University of Wales, Swansea, UK.
E-mail: J.Thompson@swansea.ac.uk J.S.Mason@swansea.ac.uk

1. ABSTRACT

A large proportion of current speech related research is founded upon databases collected under controlled conditions. The manner in which they are collected is likely induce variations due to anxiety, and analysis of a typical speech database is presented supporting this. Trends in speaker identification (SI) error rates correlate with postulated trends in anxiety levels along the time course of data collection. Further analysis using estimated glottal waveforms shows a speaker that causes high SI errors to have speech that exhibits characteristics of higher stress at the initial phase of database recording.

2. INTRODUCTION

Speech variability, both within speaker (intra) and between speakers (inter), is a topic of major interest in speech research. Much effort has been directed towards the study of variation across speakers (inter-variation), particularly with regard to speaker adaption in speech recognition systems. Significantly less research has been devoted to the study of within speaker variation (intra-variation), although it has been identified recently [3] as a topic needing a much greater understanding if speech recognition products are to gain a more widespread acceptance.

Intra-variation may be attributed to factors affecting the speaker, such as: mood/style/emotion, state of health, talking rate, prosodic context, social context, physical/mental task, background (e.g. noise level).

A large proportion of speech related research is founded upon databases recorded under controlled conditions. These controls may be aimed at certain aspects of intra-variation e.g. background, physical/mental task, textual variations, but a potentially important source which may also be present is emotional stress (anxiety) due to the unfamiliar, if not alien, situation of the recording environment. It is the aim of this paper to show: that some speakers are anxious (emotionally stressed) during database recording; the level of this anxiety changes as a function of time, see the hypothesised time profile in Figure 1; and that anxiety changes are reflected in the speech waveform as an additional uncontrolled source of intra-variation.

2.1 Effects of intra and inter speaker variation

In speech and speaker recognition systems, the amount of inter- and intra-variations present in the speech can significantly effect error rates. Speaker recognition uses the inter-speaker variability as the distinguishing factor, while the intra-speaker variability causes performance degradations [16]. In speech recognition, the situation is slightly different, with the more dominant inter-speaker variation being viewed as a major problem. In fact, some aspects of intra-variation such as phonetic and prosodic context may under some circumstances be considered useful. In both tasks however, the presence of anxiety in some speakers is likely to degrade performance, hence the results shown in this paper have relevance to both fields. In the context of database experiments, it is worth noting that performance degradation is likely to occur primarily when the anxiety level changes, giving rise to increased intra-variation. Constant style present in both training and testing are less likely to be problematic.

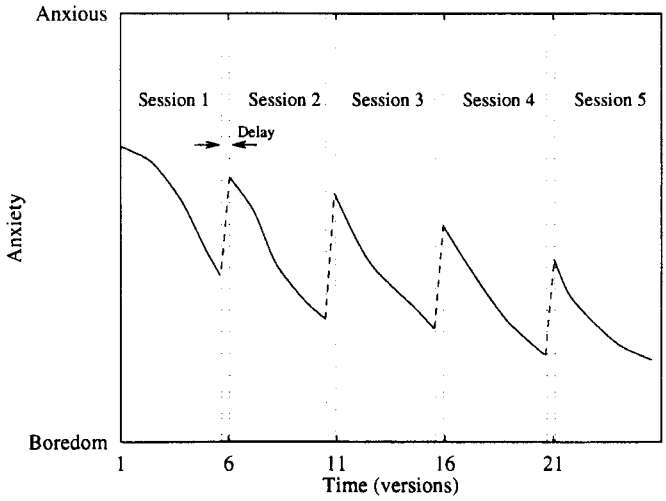

Fig. 1. Hypothesized anxiety/boredom verses recording time for a multi-session database. The anxiety in general decreases the more the equipment is used, but varies significantly within each session peaking at the beginning.

2.2 Presence of Anxiety

The task of talking to a machine, or worse still a camera is known to induce anxiety [12] [8]. This is likely to decrease over time, due to familiarisation, and may ultimately turn to boredom. A hypothetical profile reflecting these changes in a multi-session database is shown in Figure 1. Anxiety is plotted against time, in this case for 5 individual recording sessions each one separated by a delay of days, weeks or longer. It is postulated that anxiety levels at the beginning of each session are high, but fall

away during the session. This gives rise to short term and long term downward trends due to familiarisation.

The anxiety scale in Figure 1 is likely to be highly speaker, environment, and task dependent. For example the introduction of a camera is known to be a significant factor, and is employed in simulated public speaking experiments to induce anxiety [12]. The hypothesized decay of anxiety over time, and the lack of detail concerning the time between sessions in Figure 1 may be indirectly supported by the speaker recognition experiments of Furui [7] and Naik [14]. Furui suggests that data collected over a 3 month period is required to capture a speaker's intra-variations, while Naik suggests that it requires only 3 weeks. Although there seems to be a large discrepancy, the difference can probably be accounted for by Naik having much less time between recording sessions, and the size of this delay, above a certain threshold, being relatively unimportant. Both acknowledge that intra-variations occur over time, as Figure 1 suggests, however, neither offers any further explanation as to their cause, but anxiety could be one factor.

3. REFLECTIONS OF ANXIETY IN SPEECH

The presence of stress in speech is known to cause several noticeable effects. The work of Paul *et al.* [15] in the context of high noise environments, and high workload situations suggest parameters which undergo noticeable changes in the presence of speaker stress. These include:

- increased fundamental frequency (F0);
- increased frequency and amplitude of the first formant (F1);
- overall spectral tilt, speech level, and timing changes;

Others, including the early work of Millar [13] and the more recent work of Cummings [4], [5] suggest the glottal waveform as a means of identifying and characterising speaker stress. Although this is not a parameter Paul *et al.* cite, nonetheless changes in the glottal waveshape will cause changes in the spectral tilt, due to the glottal waveform contributing mainly to the wider bandwidth components of the speech spectrum [1].

The overall effect of these parameter variations can be to significantly degrade the performance of speech/speaker recognition systems, especially when not trained and tested under matched conditions, and stress levels [2]. Techniques developed for making these recognition systems robust to speaker stress include multi-style training [11], and cepstral domain compensation [2].

4. EVIDENCE OF ANXIETY IN SPEECH DATABASES

The primary purpose of the work presented in this paper is to assess the speaking style changes in databases captured under typical controlled conditions. The Millar database of British Telecom is used in subsequent experiments, as it contains a relatively large number of repetitions per person, namely 25 versions of the digit set captured over 5 sessions with typically 2 weeks separating each session, as illustrated in Figure 1;

This database was captured with speaker verification experiments in mind, not speaking style studies. However, one reason for the multiple sessions spanning several months, is to capture fully the intra-speaker variations mentioned above [7], [14].

Classification of the glottal waveform is used as the technique for detecting stress. We chose this approach for two main reasons: first, changes in the spectral tilt, hence the glottal waveform, have been noted as one of the most sensitive parameters to change under speaker stress [2], [13]; and second, analysis of the glottal waveform has already been used successfully to classify speaker stress levels [4].

Two distinct sets of experiments are considered. In the first, standard speaker identification (SI) experiments are conducted using different sets of test and training data with the objective of testing the hypothesis of style change illustrated by the saw-tooth waveform in Fig. 1. The SI recogniser uses 14th order MFCC features, and a nearest neighbour classifier, with VQ codebook models.

In the second the glottal waveform is estimated directly from the speech and analysed in an attempt to parameterise anxiety, with the immediate goal of pre-detection and a longer goal of quantifying levels.

5. RECOGNITION EXPERIMENTS

If the profile of Figure 1 crudely reflects changes in speaker anxiety during database recording then these changes if significant are likely to influence recognition accuracy. Two simple experiments were designed to test this conjecture, using the SI recogniser described above.

In the saw-tooth waveform of Figure 1, although there is a general fall across sessions, it is predicted that the most noticeable and rapid drop in the anxiety level will be within each session. Therefore, the first versions of each session, numbered 1, 6, 11, 16, and 21, are predicted to reflect a speaking style of much higher anxiety than the last versions numbered 5, 10, 15, 20, and 25. The two experiments were designed to test this hypothesis.

The recogniser is trained on version 1 and 5 separate tests are conducted:

- test 1: versions 6, 11, 16, 21,
- test 2: versions 7, 12, 17, 22,
- test 3: versions 8, 13, 18, 23,
- test 4: versions 9, 14, 19, 24,
- test 5: versions 10, 15, 20, 25,

reflecting the hypothesized high-to-low anxiety profiles. If the differences in style from the training version (1) to these test sets increase from test 1 through to test 5, then recognition errors should also increase. This is based on the assumption that the first version of each session is at a higher anxiety level than the last version, and as the recogniser is trained on version 1 of session 1, the SI error rate for test 1 should be the lowest. The results for the other tests should gradually degrade, with test 5 giving the highest errors. Figure 2 shows the actual results obtained, and as can be seen, follow the predicted pattern. This supports the hypothesis that there are more similarities with version 1 and the the first version of the remaining sessions, than with the other test set versions.

In the second experiment, the complementary case is used: training is on the very last version (25), corresponding to the lowest point of the profile in Fig. 1, and testing is again on groups from the 4 other sessions. The test

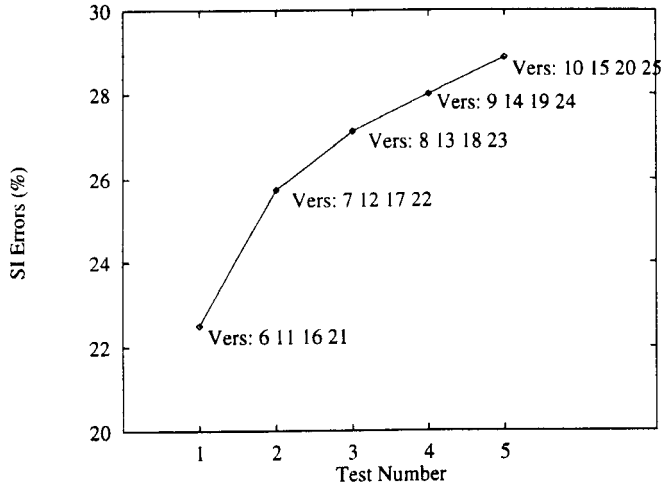

Fig. 2. SI errors (%) for 5 tests as specified, training is on version 1. Test Number corresponds to within-session version count e.g. versions 7, 12, 17, 22 used in test 2 are 2nd versions in sessions 2, 3, 4, and 5 respectively.

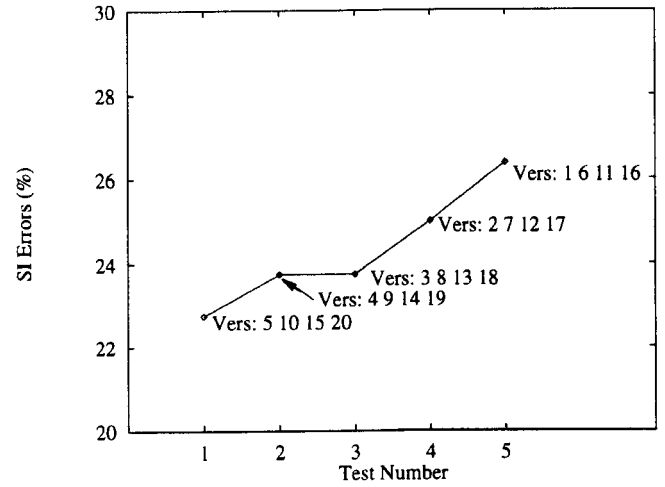

Fig. 3. As per Figure 2, but training on version 25

data is for test 1 is versions 5, 10, 15 and 20; for test 2 it is versions 4, 9, 14, and 19; etc. The results are shown in Figure 3, and show very similar trends to those from the first experiment.

These experimental results suggest the existence of short term changes in the speaking style (on average across speakers) within the sessions. They also demonstrate correlations between versions across sessions, supporting the hypothesis of the saw-tooth in Figure 1.

6. GLOTTAL PULSE ANALYSIS

In order to explain in more detail the trends in the results above, and to add evidence to the hypothesis that the intra-variations are due to some form of stress, we examine the glottal waveform. This can be used as the basis for a technique involving classification of the es-

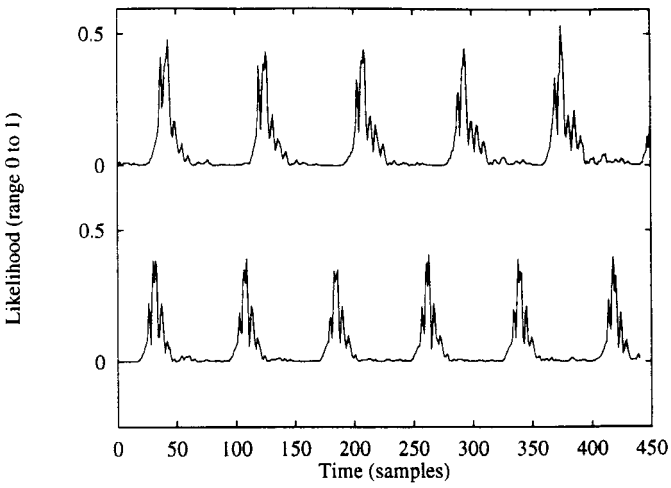

Fig. 4. Glottal excitation of speaker 1, taken from phoneme /w/ in the word 1. The top profile shows the estimated glottal excitation taken from the first version, and the bottom profile shows the estimated glottal excitation taken from the last version in the 25 version database.

timated glottal parameters extracted directly from the speech waveform.

The estimated glottal waveform is extracted using a weighted least squares lattice (WLSL) filter [6]. This is an exact least squares adaptive algorithm that has a very fast convergence. One of the filter outputs is the likelihood variable, which can be used as a measure of the 'unexpectedness' of the recent data. This rapidly tracks fast changes in the statistics of the observed data, and providing the window is made sufficiently short, can be considered as an estimate of the glottal waveform [10]. It has previously been used in a pitch detector [10], and to de-weight the effect of the glottal excitation in LPC analysis for improved format frequency estimation [9]. Here the WLSL is used in preference to more conventional inverse glottal filter techniques [17] because it processes data sample-by-sample rather than on a frame basis, with important implications for high pitch speech.

Detailed analysis of the results in the above SI experiments, show a number of speakers contribute very few recognition errors, while a group of others contribute a much larger number. In previous work, we have shown that there is a degree of correlation between recognition errors, and intra-variation [16], hence it should be possible to see whether speakers that cause high recognition errors have a higher variation in anxiety during database collection.

Figures 4 and 5 show examples of the estimated glottal waveforms extracted from two speakers, speaker 1 (male), and speaker 5 (female), as examples from the low and high error groups respectively. The speech used in both examples is for the phoneme /w/ extracted from the utterance one. Each Figure shows 2 waveforms, the top one being from the first version, and the bottom one being from the last version of the database.

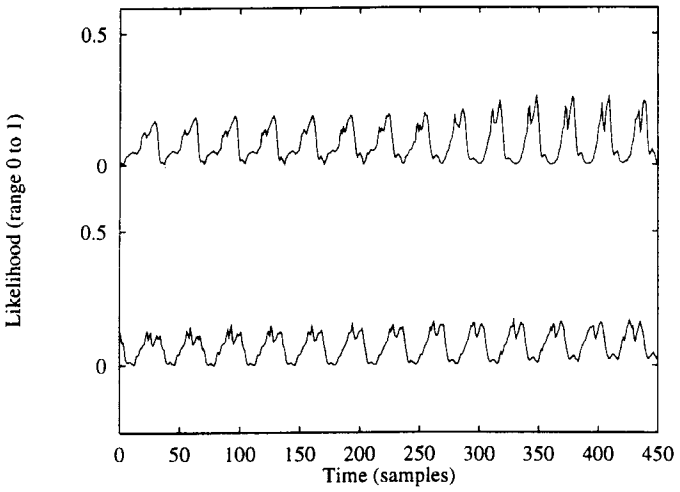

Fig. 5. Glottal excitation of speaker 5, taken from phoneme /w/ in the word 1. The top profile shows the estimated glottal excitation taken from the first version, and the bottom profile shows the estimated glottal excitation taken from the last version in the 25 version database.

The waveforms in Figures 4 and Figure 5 show significant differences between speakers, but differ much less within speakers. Characteristics of the glottal waveform from normal speech include:

> 'a slope of opening that is slower than the slope of closing.' *Cummings* [4]

Changes noted for increasing levels of speaker stress include:

> 'a marked increase in the rate of closure.' *Millar* [13]

In the waveforms of speaker 5 (Figure 5), the lower profile does conform to the characteristics stated above of normal speech. The top profile, particularly between samples 300 to 450, when considered against the work of Millar [13] does show increased stress. Also the pitch frequency is slightly higher in the top profile, something Paul *et al.* [15] identified as being a factor of stress.

Work on the classification of the estimated glottal waveforms, using the parameters identified by Cummings [4], is continuing with the explicit goal of identifying speakers with stress characteristics in their speech.

7. CONCLUSIONS

In this paper SI recognition experimental results are presented, which support the existence intra-variances in the speech that correlate well with our hypothesized anxiety levels during database collection. These demonstrate time, and in particular session variations, as postulated by the saw-tooth profile in Figure 1.

Glottal estimation from a weighted least squares lattice filter is being used to provide further evidence style changes throughout the database collection. The goal is to quantify inter- and intra-speaker anxiety variations pertaining to audio and visual speech databases.

8. REFERENCES

[1] K. T. Assaleh and R. J. Mammone. New LP-derived features for speaker identification. *IEEE Trans. Speech and Audio Processing*, 2:630–638, 1994.

[2] Y. Chen. Cepstral domain stress compensation for robust speech recognition. *Proc. ICASSP-87*, pages 717–720, 1987.

[3] R. Cole, L. Hirschman, L. Atlas, M. Beckman, A. Biermann, M. Bush, M. Clements, J. Cohen, O. Garcia, B. Hanson, H. Hermansky, S. Levinson, K. McKeown, N. Morgan, D. G. Novick, M. Ostendorf, S. Oviatt, P. Price, H. Silverman, J. Spitz, A. Waibel, C. Weinstein, S. Zahorian, and V. Zue. The challenge of spoken language systems: research directions for the nineties. *IEEE Trans. on Speech and Audio Processing, Vol. 3*, pages 1–21, January 1995.

[4] K. E. Cummings. Application, synthesis, and recognition of stressed speech. *Ph.D. Thesis, Georgia Institute of Technology*, 1992.

[5] K. E. Cummings and M. A. Clements. Application of the analysis of glottal excitation of stressed speech to speaking style modification. *Proc. ICASSP-93*, II:207–210, 1993.

[6] B. Freidlander. Lattice filters for adaptive processing. *Proc. IEEE, Vol. 70, No. 8*, pages 829–867, August 1982.

[7] S. Furui. Research on individuality features in speech waves and automatic speaker recognition techniques. *Speech Communication 5*, pages 183–197, 1986.

[8] F. S. Guimãraes, A. W. Zuardi, and F. G. Graeff. Effect of chlorimipramine and maprotiline on experimental anxiety in humans. *J. Psychopharmacology*, 3:184–192, 1987.

[9] X. Huang, G. Duncan, and M. A. Jack. Formant Estimation System Based on Least Squares Lattice Filters. *Proc. IEE, Vol. 135, Pt. F, No. 6*, pages 539–546, December 1988.

[10] D. T. L. Lee and M. Morf. A novel innovations based time-domain pitch detector. *Proc. ICASSP-80*, pages 40–44, 1980.

[11] R. P. Lippmann, E. A. Martin, and D. B. Paul. Multi-style training for robust isolated-word speech recognition. *Proc. ICASSP-87*, pages 705–708, 1987.

[12] D. M. McNair, L. M. Frankenthaler, T. Czerlinsky, T. W. White, S. Sasson, and S. Fisher. Simulated public speaking as a model of clinical anxiety. *Psychopharmacology*, 77:7–10, 1982.

[13] R. L. Miller. Nature of the vocal cord wave. *J. Acoust. Soc. Am, Vol. 31*, pages 667–677, June 1959.

[14] J. Naik. Speaker verification over the telephone network: databases, algorithms, and performance assessment. *Proc. ESCA workshop on automatic speaker recognition, identification and verification*, pages 31–38, 1994.

[15] D. B. Paul, R. P. Lippmann, Y. Chen, and C. J. Weinstein. Robust HMM-based techniques for recognition of speech produced under stress and in noise. *Proc. Speech Tech 86*, pages 241–249, 1986.

[16] J. Thompson and J. S. Mason. The pre-detection of error-prone class members at the enrollment stage of speaker recognition systems. *Proc. ESCA workshop on automatic speaker recognition, identification and verification*, pages 127–130, 1994.

[17] D. Wong and J. Markel. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. ASSP-27*, pages 350–355, 1979.