CRITICAL PARAMETERS IN THE DEFINITION OF SPEECH RECOGNISER PERFORMANCE

W. J. Barry

Department of Phonetics and Linguistics, London, UK

ABSTRACT

Six of seven "Critical Parameters" used in speech recognition assessment in the aproach known as "Recogniser Sensitivity Analysis" (RSA) are examined. An experimental study of phonetically controlled material confirms the vocabulary dependence of two of the parameters as currently defined. Basic principles of the parameter definitions necessary to assure maximum vocabulary independence are identified, and proposals for the redefinition of the Critical Parameters are presented.

1. INTRODUCTION

There have been several reports recently [4,6,7] which address the question of speech recognition assessment from the perspectives of
1) reducing the amount of speech data needed in the test database, and 2) predicting the performance of a recogniser in the field for given speaker and operational characteristics. Both these goals should be achievable if a relatively small database can be precisely defined along a number of critical parameters which specify important dimensions of speaker variability and which are not sensitive to variation in vocabulary. Situational characteristics can be simulated by post-processing "dry" recordings. This approach is known as "Recogniser Sensitivity Analysis" (RSA) [2]. Given information on the intended speakers, a controlled test of selected items from the assessment database should allow field performance to be predicted. The vocabulary independence of the parameters is crucial to the approach, otherwise every operational lexicon would have to be covered in the database.

Two projects concerned with speech recognition assessment (UK Alvey project MMI/132 - STA, and ESPRIT project 2589 - SAM) have been examining the relationship between recognition performance and the values of the test vocabulary along the critical parameters (CP), with some confirmation that variation in recogniser performance can be explained by variance in several of the parameters [7]. However, the evidence is not unequivocal [3]. This may well be a consequence of the recognisers selected for the trials, but the results are such that a scrutiny of the parameters and their definitions is warranted. This paper firstly undertakes this scrutiny, secondly presents results of an experiment which illustrates the inherent vocabulary dependency of two of the Critical Parameters as defined at present, and thirdly suggests some modified parameter definitions which are more appropriate to the underlying rationale of speech recognition.

2. CP DEFINITION

The present CP definitions are as follows [5]:

1. *Speaking rate:* Duration of one utterance relative to the overall average of all utterances of that word.
2. *Vocal tract area:* Average VTA over the frames of one utterance

in relation to the overall average VTA for all utterances of that word.
3. *Temporal congruence:* Average DTW distance between all pairs of utterences of the same word by one speaker. The distance value applies to all utterances of a given word by that speaker; it is a measure of speaker consistency.
4. *Vocal effort:* Ratio of peak and average energy. Calculated for each utterance independently.
5. *Spectral definition:* Average (and variance of) ratio of energy < 2kHz and total energy, computed over each utterance independnetly.
6. *Fundamental frequency:* Mean F0 over each utterance independently.

A seventh parameter, vocabulary complexity, lies outside the focus of this paper.

3. TEST OF VOCABULARY INDEPENDENCE

The above definitions make it most likely, *a priori*, that "spectral definition" and "vocal effort" are strongly vocabulary dependent. Using the SAM_SPEX CP-extraction software, developed for use on the SAM PC-based workstation (SESAM) by Jutland Telephone according to the Logica algorithms [5], two sets of phonetically controlled words were analysed with respect to these two CPs:
1) 18 /hVd/ words spoken 3 times with different intonation contour by 4 speakers (2F, 2M). These were selected at random from the 10F and 8M speakers recorded in the Normative Reference database of the UK Alvey project MMI/132 (STA) [1]. These words were examined to test the effect of vowel variation on parameter values.
2) Five phonetically varied words selected from Logica's RSA recogniser test vocabulary [7] ("Aberdeen, Darlington, Manchester, Ipswich, sixteen") These were spoken 5 times each with 4 different voice qualities: modal, breathy, creaky and falsetto). With the combination of open and close vowels, and the presence or absence of consonants with fricative elements, these words tested the combined effect of changes in vowels and consonants on the parameter values.

Speaker variation was simulated in the second set by using the same, experienced speaker to record the words with the four radically different voice qualities.

4. RESULTS

The results may be summarised as follows:
1) Values in the /hVd/ words were critically dependent on the threshold setting chosen for endpointing. The 10% (of mean energy) threshold chosen as default for the tests was clearly too high to capture the tri-phonemic structure of these words. The parameters were, in effect, being calculated over the vowel portion alone. None of the words in the second set were affected by this threshold setting; it is solely a problem for words beginning with acoustically "weak" consonants (e.g. /h, f, v, T, D/- SAMPA symbols [8] are used throughout) and/or ending with such consonants, including weakly exploded stops. Reducing the threshold to 5% and 3% in two further analyses allowed the full /hVd/ word to be captured, but it is clear that this also increases the risk of including breath-noise and lip-smacks in the parameter estimation (and the recognition process) if they occur immediately prior to the actual utterance.
2) Two of the parameters, as defined at present: "spectral definition" and "energy" are dependent on the phonetic structure of the word.

A two-way ANOVA was performed on both sets of data for each parameter. For the /hVd/ words, "spectral definition" varied significantly with both speaker (F = 15.05, DF 2) and word (F= 15.06, DF 17). Energy" did not vary significantly across the words, intervowel intensity variation not being great or regular enough. For the second set of words, "spectral definition" again varied highly significantly across the words (F = 51.51, DF 4) and also reached significance for voice quality (F =

3.09, DF 3). With the "energy" parameter, there was significant interaction between voice quality and word (F = 11.54, DF 79).

The phonetic factors underlying these differences can be summarised as follows: 1. *Spectral definition* (Voice quality) (Energy < 2kHz/total energy): In the /hVd/ words, the main difference was between words containing mid to high front vowels and the others. This is due to the presence of strong higher formants, and especially the high second formant (> 2kHz) reducing the energy quotient. However, this variation was small compared to the value shifts across the words in the second set, where the presence or absence of fricatives (high frequency noise) changed values by up to 25%. 2. *Energy* (peak energy/mean energy) maximises the vocabulary effect and thus undermines speaker differences by relating peak energy (= peak energy of stressed vowel) to mean energy (which is proportionally lower, the more unvoiced consonants a word contains).

In the /hVd/ words, the values are most dependent on the personal strength of production, the intrinsic intensity of the vowels (open > close vowels) not contributing to any appreciable extent; in the second word-set, the greater complexity of the word structure contributes to a very strong interaction between differences in energy resulting from the different voice qualities and the word. For example, the consonants in "Ipswich" reduce the mean energy, and increase the quotient considerably despite the vowel /I/, which has relatively low intrinsic intensity. So, despite consistently lower quotient values for each individual word for the "breathy" than for the "modal" voice quality, the values for breathy "Ipswich" are higher than the "modal" values for the other words.

5. DISCUSSION OF DEFINITIONS

In the light of these results, it is important to consider how possible improvement in the word independence of these measures can be achieved. Firstly, for "spectral definition" and "energy" separating the main source of word-dependent variability, the voiced and voiceless portions, is an important prerequisite. For "spectral definition", only the voiced portions would be used in the calculation, reflecting the rationale behind the parameter, namely of capturing some aspect of voice quality. For "energy", mean voiceless intensity should be related to mean voiced intensity.

Secondly, for all measures, it is essential to avoid any utterance-dependent expression of peak value, mean value or variability. Speaker-dependent aspects of the parameters can only be calculated by relating the value for an individual utterance to the mean of all utterances of a particular word. In speaker-dependent recognisers, it may be sufficient to relate the parameter value for the individual utterance to the mean for all utterances of the same word/expression by the same speaker. In general, however, for speaker-independent systems the individual value has to be related to the mean of all utterances of the word/expression for all the other speakers (temporal congruence is the one exception being a speaker-dependent measure). This is already done with "speech rate" and "vocal-tract area" where the relativity of the measure was clear from the outset. Such normalisation should, however, be extended to "voice quality", "intensity" and "F0".

Thirdly, the "fundamental frequency" parameter in the form of *mean* F0 is dispensible, because it only separates male from female at present. This is information that is available independent of analysis. F0 variance, which is currently also being calculated, may be a more useful measure, since wide fluctuations in F0 could correlate with variation in recogniser performance; in Logica's recogniser tests [7], variance appears to be a more important factor than mean

F0. The measure should, however, be related to the mean to avoid confounding Hz variance and the male-female distinction. The coefficient of variance (the quotient of mean and standard deviation) is therefore suggested. However, this measure still requires normalisation by relating it to the mean of the other speakers.

The principle of taking variance rather than mean values should also be considered for "vocal tract area". Although, in its present definition, it is already normalised, it is the mean of the frame-based area coefficients which are being normalised. The use of variance would relate the parameter more closely to the recognition process. It is, after all, *variation* of signal properties during the course of a word which makes it more or less distinctive to a recogniser not its *mean* sig-nal properties. Variance in vocal tract area for a given utterance related to all speakers' variance for utterances of the same word or expression would differentiate speakers with clear and less clear articulation.

6. CONCLUSIONS

The results from the CP analysis of phonetically controlled vocabulary confirmed the vocabulary dependence of two of the Critical Parameters as currently defined for RSA. It was concluded that parameter normalisation across all utterances of a word is a fundamental pre-requisite for the vocabulary independence of Critical Parameters. Further consideration of the basic rationale underlying automatic speech recognition points to the need to redefine all parameters except "temporal congruence" to contain an expression of normalised variance.

7. REFERENCES

[1] FULLER, H., FOURCIN, A.J., GOLDSMITH, M.J. and KEENE, M. (1990): A Database of Normative Speech Recordings. *Proceedings Institute of Acoustics* 12, part 10, 1-6

[2] KNIGHT, J.A. and PECKHAM, J.B. (1984): *A Generic Model for the Assessment of Speech Input Applications.* Logica Report for RSRE.

[3] KORDI, K. (1990): *Field Trial Report.* Alvey Project MMI/132, Logica, November 1990

[4] PECKHAM, J., THOMAS, T. and FRANGOULIS, E. (1989): Recogniser Sensitivity Analysis: Trial Results and Future Directions. *Proc. ESCA Workshop, Speech Input/Output Assessment and Speech Databases*, 4.2.1-7, Noodwijkerhout.

[5] THOMAS, T.J. (1988): *Algorithms Used for Parameter Extraction for Recogniser Sensitivity Analysis.* Project Report. Alvey Project MMI/132, Logica, July 1988

[6] THOMAS, T.J. (1989): A Determination of the Sensitivity of Speech Recognisers to Speaker Variability. *Proc. ICASSP*, S10b.8, 544-547

[7] THOMAS, T.J. (1990): *The Sensitivity of a Speech Recogniser to Speaker Variability.* Project Report. Alvey Project MMI/132, Logica, Cambridge, September 1990

[8] WELLS, J.C. (1988): Computer Coded Phonetic Transcription. *J. International Phonetic Association* 17 no. 2, 94-114.