Proceedings of The Institute of Acoustics

THE ANALYSIS AND CLASSIFICATION OF F0 CONTOURS

D. Lindsay

Department of Communication and Neuroscience,
University of Keele, Keele, Staffordshire, England.

INTRODUCTION

There have been many proposals for prosodic components in automatic speech recognition systems [1,2]. Lea [3] suggests the use of a prosodics-only recogniser to investigate various features of intonation. However, that was in connection with the interpretation of intonation as the signalling of juncture phenomena, especially coordination and subordination. The work presented here takes this approach further and analyses the intonational nucleus in terms of a a limited number of types according to a theory of the phonology of British English intonation. The nucleus is typically identified by the prominence of a syllable, the range of associated pitch movement, or the informational focus of an utterance. The intonation nucleus has been described by Halliday as the central, obligatory component of his theory of British English intonation [4]; the nucleus is described as being one of five (simple) contour types when the pitch movement is associated with a single syllable.

There is evidence that Halliday's five tone classification is an accurate model of nuclear intonation from recent work on the resynthesis of speech [5] (using variants of the five tones) and from a direct comparison with the fall/rise, statement/question model of nuclear intonation [6]. The theory provides a model of the more central phonological functions of intonation and provides a basis for extending the operation of an automatic intonation analyser. It would be interesting to determine how far that recogniser can, in fact, perform the classification of nuclear tones in terms of those five simple tones of English.

SINGLE TEMPLATE F0 ANALYSIS

The classification of the sample F0 contours and the training of the classification system is performed by a modification of a system developed for the automatic analysis of F0 contours [7]. F0 templates are chosen by a training run of the classifier on a subset of the sample population. Non-linear time normalisation (using dynamic time warping) is used to achieve an optimal match between these F0 contour templates and stretches of preprocessed speech. This process was carried out for speaker dependent and independent classification, F0 estimation confidence weighted and unweighted, and for a range of DP adjustment window lengths.

The preprocesed speech data is obtained by direct analogue-to-digital conversion, sampling at 5 kHz, using an antialiasing filter, and with a word size of 12 bits. The signals were taken from the Lx output from a Laryngograph. The processed speech representation consists of a sequence of parameter sets; each parameter set is a transformation of a 25.6 ms rectangular time window of speech; the frame rate of this window is 100 Hz. The parameter set consists of two measures: a logarithmic transform of fundamental frequency

general construction rules can be found. These rules have the same form as phonological rules for connected speech (although the "symbols" may not have a traditional phonetic interpretation) and can be used to convert a network of standard word pronunciations to a network of acoustic segments as performed in the HARPY system. This integrated syntactic, phonological and acoustic knowledge representation is an effective and efficient structure for connected-word recognition.

CONCLUSION

The key to the link between the phonetic form and the acoustic form presented in this paper is that a network representation of acoustic similarities found by Moore's algorithm can be mapped to a network constructed from context-sensitive rules operating on phonetic symbols. Until acoustic models of linguistic units (words/phonemes/phones) become more refined, the explanation of a phrase as a concatenation of acoustic segments may be useful for incorporating phonological knowledge into speech recognition systems. The approach is not another new architecture for speech understanding but an extension of proven connected-word recognition procedures.

Finally, there are still questions to be addressed:

- Which test words are required for acoustic segmentation ?
- How many segments are required to reach the fidelity criteria ?
- Which connected speech phenomena are worth modelling ?
- How might the segmentation process be automated ?

Investigation of these questions is taking place at UCL.

REFERENCES

[1] B.T. Oshika, V.W. Zue, R.V. Weeks, H. Neu, J. Aurbach, "The role of phonological rules in speech understanding research", IEEE Trans. ASSP 23 (1975) p104.

[2] C.J. Weinstein, S.S. McCandless, L.F. Mondshein, V.W. Zue, "A system for acoustic-phonetic analysis of continuous speech", IEEE Trans. ASSP 23 (1975) p24.

[3] B. Lowerre, D.R. Reddy, "The Harpy speech understanding system", in Trends in Speech Recognition ed W. Lea, Prentice Hall 1980.

[4] J.S. Bridle, M.D. Brown, "Connected-word recognition using whole-word templates", Institute of Acoustics Conference 1979.

[5] R.K. Moore, P. Beardsley, M.J. Russell, M.J. Tomlinson, "Towards an integrated discriminative network for automatic speech recognition", Institute of Acoustics Conference 1982.

This research is funded under MOD University Research Agreement 2047/0104/RSRE.

THE ANALYSIS AND CLASSIFICATION OF F0 CONTOURS

$$g(I,J)/N$$

where N = I + J.

SAMPLE POPULATION

The two databases used for training and classification partitioned a set of 400 speech utterances from four male British English speakers. The samples consisted of five repetitions of each of four syllables spoken in five ways. The subjects were prompted to speak with the specified intonation by asking for the samples to be spoken 'as a statement', 'as a question', 'weakly', 'with reservation', 'with emphasis'. The prompts were based on the semantic function labels given by Halliday. The labels were

1. statement
2. question
3. weak statement
4. reservation
5. emphatic statement

Gross errors in pronunciation were rejected, although the subjects produced natural sounding stretches of speech with little prompting. The carrier syllables were chosen to have continuous voicing for most of the extent of the utterance and also to be neutral and common enough to occur in informal conversation; it was decided that this would contribute towards a natural intonation being used. The syllables were

1. yes
2. no
3. mmm . .
4. well

The relevant sections of the preprocessed data were selected by an automatic routine: a smoothed copy of the CF0 measure of the parameter sets was used to decide on the region of high confidence corresponding to good estimates of F0. The smoothing was average-of-four and the threshold was 30 on the self-normalised CF0 scale ranging from 0 to 100. All estimates within the first points of exceeding and falling below the threshold were assumed to be the desired region. The unsmoothed CF0 values were used in the subsequent analysis and classification.

TRAINING

The classifier was trained by selecting a set of five reference templates, each of which minimised the sum of the intra-class distance for the respective classes using the dynamic time warping error as a measure of sample-to-sample distance. In practice, for any one sample this involved the summing of all the scores for matching with all other training samples in its classification subclass. Five reference templates were chosen in this way for the classifier each time a change was made in one of the experimental parameters. The selection for each class was on eight samples for each tone class for each of the four subjects in

the speaker dependent mode.

There was little variation over the adjustment window parameter for all the subjects, and what differences in selection exist are found on the extrema of the window range. This suggested that there was a steady-state region for r=3 to r=4 (15-20% of sample length) in respect to speaker dependent training; it could be seen from the classification experiments that there is no measurable improvement in the recognition scores for r>3. The speaker dependent training (and the classification) appears to stabilise at around the window length of 3 so it was decided to train the independent mode for that value. The selection for each class for the independent mode was on 33 samples for each of the five tone classes.

CLASSIFICATION

The classification of the test samples was achieved by calculating the DTW distance from each of the five reference samples and choosing the minimum as indicating class membership.

Speaker dependent classification. The effect of varying the adjustment window length on classification performance is given in Table 1; this is the overall performance averaged over the five tones, each score value represents the classifier performance for 60 test samples. It can be seen that there is a slight, but consistent, improvement from r=1 to r=3, thereafter levelling out. The breakdown for performance within tone categories shows similar trends, no class excepted, but with three classes, one, two and four, achieving a consistently higher score.

r	subject	
	1	2
1	79.2	76.2
2	88.1	75.0
3	89.9	76.6
4	89.9	76.6

Table 1. Effect of window length on speaker dependent scores (% correct)

Speaker independent classification. The classification scores for the five tones using the speaker independent mode of operation is given in Table 2, each score the result of 240 classifications. The tabled scores are for a window length of 3, for both CF0 weighted and unweighted DTW distance measure.

THE ANALYSIS AND CLASSIFICATION OF F0 CONTOURS

tone	weighting	
	yes	no
1	56.2	56.2
2	70.8	70.8
3	63.8	63.8
4	64.6	64.6
5	60.4	66.7

Table 2. Speaker independent recognition scores - overall (% correct)

The scores are similar, suggesting that the CF0 weight as a confidence measure on the F0 estimate is not a significant factor when the F0 estimate is consistently good. The F0 values generally had a high confidence measure for all the preprocessed speech samples in these experiments. The scores are generally a good deal lower that those for the dependent mode. A breakdown of these scores across the four speakers is given in Tables 3 (weighted) and 4 (unweighted). The scores here show less consistency, the tones from some speakers completely misclassified throughout, and the tones from the subject providing the reference template achieving higher than average classification. The overall scores for all speakers and all tones were 63.2% (weighted) and 64.4% (unweighted); that compares with the overall score for speaker dependent classification of 83.2%

subject	tone				
	1	2	3	4	5
1	100.0	41.7	63.6	91.7	75.0
2	83.3	100.0	100.0	75.0	41.6
3	41.7	66.7	25.0	41.7	83.3
4	0.0	75.0	66.7	50.0	41.6

Table 3. Speaker independent recognition scores - weighted (% correct)

subject	tone				
	1	2	3	4	5
1	100.0	41.7	63.6	91.7	75.0
2	83.3	100.0	100.0	75.0	50.0
3	41.7	66.7	25.0	41.7	83.3
4	0.0	75.0	66.7	50.0	58.3

Table 4. Speaker independent recognition scores - unweighted (% correct)

CONCLUSIONS

The experiments described here show that a prosodics-only recogniser is feasible, but only to a limited extent. Using a description of nuclear intonation derived from Halliday's theory, a classification of F0 contour types was possible, both in speaker dependent and independent operation, although the performance falls short of that achieved by current segmental recognition schemes. Work is being carried out to extend the size and range of the database, and to determine the applicability of this analysis scheme to the classification of contours other than those associated with the nuclear syllable.

ACKNOWLEDGEMENTS

This work was supported by the Joint Speech Research Unit, Cheltenham.

REFERENCES

[1] W.A. Lea, M.F. Medress, and T.E. Skinner, 'A Prosodically guided speech recognition system', IEEE Transactions ASSP-23, 30-38, (1975).

[2] J. Vaissiere, 'Speech recognition programs as models of speech perception', in The Cognitive Representation of Speech (eds. Myers, Laver, Anderson), Amsterdam: North-Holland, 443-458, (1981).

[3] W.A. Lea, 'Prosodic aids to speech recognition', in Trends in Speech Recognition (ed. Lea), New-Jersey: Prentice-Hall, 166-205, (1983).

[4] M.A.K. Halliday, The tones of English', Archivum Lingusticum, vol. 15, 1-28, (1963).

[5] N.J. Willems, 'STEP: A model of standard English intonation patterns', IPO Progress Report 18, 37-42, (1983).

[6] W.A. Ainsworth and D. Lindsay, 'Identification and discrimination of Halliday's primary tones', this volume.

[7] D. Lindsay, 'A method of describing pitch phenomena', in Investigations of the Speech Process (ed. Winkler), Bochum: Studienverlag Brockmayer, 189-210, (1983).

[8] H. Sakoe and S. Chiba, Dynamic programming algorithm optimisation for spoken word recognition', IEEE Transactions ASSP-26 (1), 37-51, (1978).