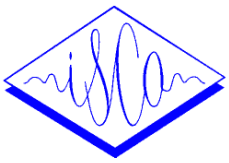

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

MODELLING OF INTONATION CONTOURS AT THE SENTENCE LEVEL USING CHMMS AND THE 1961 O'CONNOR AND ARNOLD SCHEME

U. Jensen[a], R.K. Moore[b], P. Dalsgaard[a], B. Lindberg[a]

[a]Center for Personkommunikation, Aalborg University, Denmark
[b]Speech Research Unit, Defence Research Agency, United Kingdom

ABSTRACT

This paper describes work on the recognition of British English intonation contours at the sentence level, using continuous density hidden Markov models (CHMMs) and the 1961 O'Connor and Arnold scheme for describing intonation contours. The scheme is presented briefly and the model optimisation procedure is described. Experimental results are presented for four tasks: transcription, type of nuclear tone, position of the voiced part of the accents and the nucleus (or principal accent). Taking into consideration the small training and test material, the results show a fairly high accuracy in all four tasks.

10.21437/Eurospeech.1993-185

1. INTRODUCTION

Prosodic information provides a rich yet mainly untapped source of information and constraint for speech based interactive applications. As capabilities of automatic speech recognition and speech generation expand to embrace larger vocabularies and more complex syntactic and semantic constructions, so a greater understanding will be required of the *suprasegmental linguistic structure*. An important aspect of suprasegmental structure is *intonation* which constitutes the tune (or melody) in spoken language. Intonation is closely related to *fundamental frequency* (F_0) and *pitch* as well as being influenced by the pattern of stressed and unstressed syllables [1].

Little attention has been given to the modelling of intonation contours in automatic speech recognition, and this is probably due, in part, to the more general difficulty of describing and classifying such contours, and assigning appropriate (semantic) meaning to them. However, intonation contours are clearly important, as evidenced by the vast amount of published literature in the area e.g. [1,3,4]. On the other hand, it has yet to be established to what extent intonational information might be valuable in automatic speech recognition and understanding. Also, before any significant experimentation in this area can be initiated, reliable systems for extracting intonational information have to be developed.

The scope of this work is to develop a system for modelling intonation contours at the sentence level, in British English, using *hidden Markov modelling* (HMM) techniques. The appropriateness of using HMMs for modelling word-level prosodic and intonational information has already been addressed in [5,6,7]. The work presented in this paper is an extension of these investigations to cover complete sentences.

Three main issues arise in the development of an HMM based system for recognising intonation contours at the sentence level. These are: **a**) to define a set of descriptive units for describing intonation contours at the sentence level including a grammar, **b**) to select a suitable set of features and to investigate methods for their extraction and **c**) to optimise the topology of the models taking into account the defined units and the behaviour of the features within voiced and unvoiced segments.

These three issues are addressed in sections 2 and 3 together with a brief description of the O'Connor scheme, the data used in the experiments, the feature extraction process and the problems involved in modelling F_0-contours. The results of the experiments are presented in section 4 and a conclusion is given in section 5.

2. INTONATION

The traditional way of describing intonation contours in British English is based on *intonation groups* (or similar such units) together with a syllable which is designated to be the centre of the intonation group (the *nucleus*) [3,8,9,10]. However, in the four works referenced, schemes of varying complexity are suggested for describing the contours within each intonation group; [3] being the most complex and [8] the least. In these preliminary experiments the least complex of the schemes is used.

2.1 The O'Connor scheme

In the 1961 O'Connor and Arnold scheme, tunes are described in relation to intonation groups. Disregarding the phenomenon of compound tunes, each intonation group consists of three *intonation segments* which may or may not be present: *pre-head*, *head* and *tail*, and one intonation segment which is always present: *nucleus*. The pre-head consists of any syllable before the first accented syllable. The head begins with the first accented syllable (before the nucleus) and ends with the syllable immediately preceding the nucleus. Finally, the tail consists of all syllables following the nuclear syllable. An example is given below; accented syllables are written in capital letters.

he WANTS to be ABsolutely SURE about it

pre-head head nucleus tail

The perceived pitch contour in each intonation segment is described by means of a finite number of stylised patterns (or symbols) which are considered suitable for being modelled by HMMs. The O'Connor scheme consists of 15 such symbols. However, four of the symbols represent minor intonational events; hence they are discarded in the present system. The 11 descriptive units distinguished are: two pre-head symbols (high '‾' and low ' '), three head symbols (low '‖', stepping '↗' and sliding '↝') and six nuclear symbols (low fall '↘', high fall '↘', rise-fall '^', low rise '↗', high rise '↗' and fall-rise 'v'). The signs presented in quotes are the signs used for transcribing utterances. A likely transcription of the example given previously would be:

he ↗wants to be ↗absolutely ↘sure about it

Hence, the sentence consists of a low pre-head, a stepping head with two steps and a low fall.

The 11 symbols do not represent patterns of equal complexity; the high pre-head is the least complex, and fall-rises and rise-falls are the most complex. Also there exists a high degree of variability in the patterning associated with a given symbol. For example, the rise-fall can be associated with four basic patterns (see Table 1), each being produced by different combinations of pitch glides and jumps.

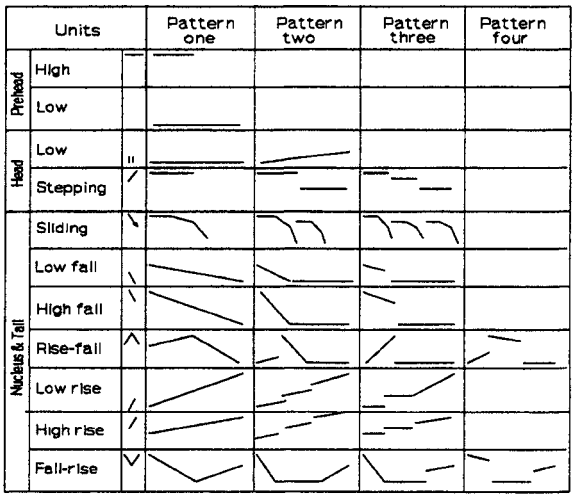

Table 1. *Intonation segment characteristics. For the stepping and sliding head, a one, two and three step/slide is shown*

2.2 Stress levels

In relating sentence stress to the reduced O'Connor scheme and to intonation groups, it is possible to define four levels of stress/accent within each intonation group. This is suitable since stress in connected speech occurs in varying degrees of prominence. The four levels of stress/accent distinguished are: *primary stress, secondary stress, tertiary stress* and *unstressed* [1]:

PRIMARY STRESS	nuclear symbols
SECONDARY STRESS	head symbols
TERTIARY STRESS	unmarked
UNSTRESSED	pre-head symbols + unmarked

The syllable carrying primary stress, or principal accent, is the most prominent syllable in the intonation group. Hence, it is always marked by one of the nuclear symbols in the O'Connor scheme. Furthermore, all syllables marked with a head symbol are said to carry secondary stress, or subsidiary accent. All other syllables either carry tertiary stress or are unstressed. In the example given, 'sure' carries primary stress, 'wants' and 'ab' carry secondary stress while the rest of the syllables are non-accented.

3. GENERATION OF MODELS

The symbols in the reduced O'Connor scheme are represented by strictly sequential left-right CHMMs, where each output pdf is composed of a mixture of multivariate Gaussian pdfs with diagonal covariance matrices. Models are trained using the SIRtrain software package [11] which is based on Baum-Welch reestimation.

3.1 Speech corpus

The speech corpus used has been specially recorded with intonational aspects in mind. The data consists of pre-marked sentences, of the form given in the example taken from [8]. A phonetician, who is familiar with the O'Connor scheme, read the sentences aloud according to the markings given. After the recording, the phonetician listened to the material and re-marked those utterances which did not fit the transcription given in the script. This secured a consistent relation between the markings and the actual contours. For a description of how the data was recorded, see [12].

After the recordings, symbol boundaries were marked and the data split into a training and a test corpus. The training corpus contained 325 symbols in all; the stepping head being the most common with 81 occurrences and the high rise the least with 11 occurrences. For the test corpus the numbers were 170, 41 and 6, respectively.

3.2 Feature extraction

The principal acoustic correlates to intonation are *fundamental frequency, energy* and *syllable duration* [1]. However, F_0 is the feature most closely related to intonation. Hence, to limit the scope of these investigations, only F_0 and F_0 derived features are used. Observation vectors containing F_0, the first time difference (ΔF_0) and the second time difference ($\Delta^2 F_0$) are calculated every 10 ms using the following expressions:

$$\Delta F_0\,[n] = F_0\,[n] - F_0\,[n-1]$$
$$\Delta^2 F_0\,[n] = \Delta F_0\,[n] - \Delta F_0\,[n-1]$$

3.2.1 *F_0-estimation*

Two F_0-algorithms have been considered in this project: the simplified inverse filter tracking algorithm (SIFT) [13] and an algorithm based on [14] which is implemented in the speech analysis program *waves* [15]. Both algorithms have difficulties in estimating F_0 over the entire F_0 interval spanned by the data (approx. 100-500 Hz). However, the algorithm implemented in *waves* is used in the system reported.

3.2.2 *Post extraction processing*

The output of an F_0 algorithm is, in general, a sequence of zero estimates (unvoiced segments, pauses and dropouts) and non-zero estimates (voiced segments). Additionally, the non-zero F_0 estimates may be divided into two categories: estimates which are either halved or doubled due to algorithmic errors and all other non-zero estimates which may be regarded as *valid* estimates.

Due to the nature of the data and the algorithmic errors which occur, the raw observation vectors have to be processed prior to modelling. Each time F_0 changes from or to a zero estimate, a halved/doubled estimate or a valid estimate, the current value of ΔF_0 and the current and the following values of $\Delta^2 F_0$ value, do not carry any information. In the present system, these boundary values are set to zero.

3.3 The two fundamental problems

Three kinds of observation vectors have to be handled appropriately in modelling intonation contours: the valid vectors which are those mainly carrying the intonational information, the vectors based on halved/doubled F_0 estimates, and zero vectors. Two fundamental problems arise from the presence of non-valid observation vectors; those which have to be eliminated prior to the modelling and those which have to be handled explicitly in the CHMMs.

To handle the zero vectors, three methods have been investigated: **a)** remove zero vectors from observations prior to the modelling, **b)** model zero vectors explicitly in the CHMMs by adding an extra pdf to each mixture and **c)** replace each zero vector by a random vector generated from a pdf with a large variance and then model the

random vectors explicitly in the CHMMs. All three methods give similar system performances. However, solution a) is preferred since it introduces a 60% data reduction and requires one less pdf in each mixture. Furthermore, this approach is consistent with the perception of pitch contours, since *'an ear listening for an overall pitch pattern learns to ignore gaps in voicing'* [1].

Model	Tying	N	M
High pre-head	yes	2	9
Low pre-head	yes	3	9
Low head (fall)	no	3	2
Low head (rise)	no	3	2
Stepping head	no	3	3
Sliding head	no	3	3
Low fall	yes	4	9
High fall	yes	5	9
Rise-fall	yes	6	9
Low rise	yes	4	9
High rise	yes	5	9
Fall-rise	yes	6	9

Table 2. *Optimised model topology (N = number of states, M = number of Gaussians in each mixture)*

The problem of F_0-halving/F_0-doubling is handled using within-state tying of means (μ) and variances (U) of individual Gaussian distributions. They are tied together in sets of three in order to make up a pdf specifically designed to model the behaviour of the F_0-algorithm. In each *pdf-set*, one pdf models the halved estimates (pdf_{half}), one models the doubled estimates (pdf_{doub}) and one models the valid estimates (pdf_{valid}). Tying ensures that $\mu_{half} = \frac{1}{2}\mu_{valid}$, $U_{half} = \frac{1}{4}U_{valid}$, $\mu_{doub} = 2\mu_{valid}$, $U_{doub} = 4U_{valid}$ after the reestimation. In general, parameter tying does not alter convergence properties in the reestimation process [16]. However, in this special case proper convergence is only ensured if pdf_{valid} 'really' models the valid estimates, pdf_{half} 'really' models the halved estimates and pdf_{doub} 'really' models the doubled estimates [17].

3.4 Optimised model topologies

One model is trained for each of the symbols in the reduced O'Connor scheme, except for the low head which is represented by two models. There is also one model for low heads preceding low falls, high falls and fall-rises and one for low heads preceding low rises, high rises and rise-falls. Each symbol is modelled by a CHMM of different complexity. Lower complexity models are used for modelling pre-heads and heads and high complexity models for modelling the nuclear tones (see Table 2). Furthermore, parameter tying is not used in the head models since occasionally there is an octave difference between valid F_0 estimates. Consequently, proper convergence properties can not be ensured.

4. EXPERIMENTS

The system has been tested on four tasks, using a recogniser [18] which is based on the token passing algorithm [19]. The four tasks are: transcription, type of nucleus, position of the voiced part of accents and nuclei. All test sentences contained one intonation group only.

4.1 Transcription

The transcription task is a conventional recognition experiment in which the system transcribes test utterances and is then assessed by comparing its output with the true transcription. Results are presented in terms of percent symbol accuracy which is computed using dynamic programming to align the true transcription of the test data with the output of the recogniser:

$$\%\ \text{sym accuracy} = \frac{T - S - D - I}{T} \cdot 100 \qquad (1)$$

where T is the number of symbols in the test set, S is the number of symbols recognised incorrectly, D is the number of symbols deleted and I is the number of symbols inserted. In the alignment process penalties for substitutions, deletions and insertions were all set to 1.

Allowing for substitutions between the two low-head models, the results in Table 3 show a performance of 67.1% symbol accuracy. 122 out of 170 symbols are correctly recognised while the number of deletions and insertions are 17 and 8 respectively.

Segment	% symbol accuracy	95% confidence
All	67.1	59.7 - 73.7
Pre-head	70.7	55.5 - 82.4
Head	61.2	49.2 - 72.0
Nucleus	71.0	58.8 - 80.8

Table 3. *Transcription task*

4.2 Type of nucleus

The task of nuclear tone recognition can be seen as a simplification of the transcription task. Only the identity of the nuclear tone is considered; whether the system recognises the pre-nuclear patterns correctly or not has no significance. In this condition, changing the original system by reducing the number of pdfs to one in the mixtures of the pre-head and head models, increases the performance from 71.0 to 77.4% symbol accuracy.

4.3 Position of accents

Determining the position of the accented syllables is equivalent to determining the position of the head and the nuclear symbols (see section 2.2). Hence, the system described automatically estimates the position of the voiced part of accents. Performance on this task is presented as % position accuracy, which is calculated for the optimal alignment as follows:

$$\%\ \text{position accuracy}_X = \frac{C_X - I}{N} \cdot 100 \qquad (2)$$

where C_X is the number of accents for which the position of the voiced part is correctly estimated with a maximum deviation of X ms, I is the number of insertions and N the number of accents in the test data.

In Figure 1, % position accuracy is plotted against X for both the whole test data and the subset of 20 utterances which are correctly transcribed. Far better performance is achieved for the correctly transcribed data e.g. % position accuracy$_{20}$ is only 33.3 for the data as a whole, while it is 64.0 for the correctly recognised utterances.

4.4 Position of nucleus

Each utterance consists of one intonation group only. Hence the last

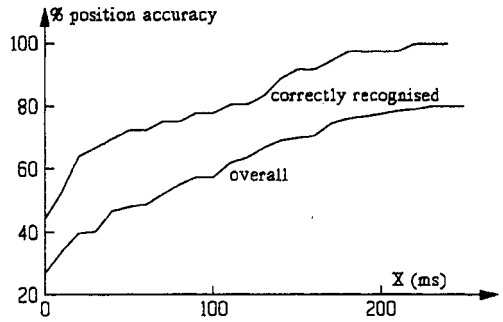

Figure 1. *Estimation of accent position*

accent estimated by the system is always the nucleus and therefore performance may be computed using a slightly revised version of formulae 2 where $I = 0$.

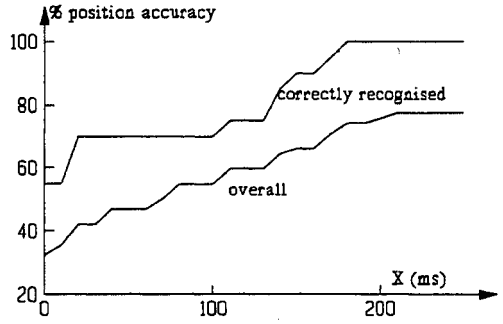

Figure 2. Estimation of nucleus position

In Figure 2, % position accuracy is plotted against X for both the whole test data and the subset of utterances which are correctly transcribed. For these % position accuracy is 70.0 while it is 42.0 for the whole data set.

5. CONCLUSION

A system for recognising intonation contours at the sentence level using CHMMs and the O'Connor scheme for describing the contours has been developed. It has been tested on both transcription and positioning tasks showing a fairly high accuracy on both. However, the present system is limited in a number of ways. Further investigations will concern the removal of these limitations e.g. speaker independence, using speech material which is not recorded with intonation in mind and a less restricted grammar which does not require that each utterance is one intonation group. Additionally, it is important to develop more robust F_0-algorithms and investigate the relevance of other features to modelling of intonation contours such as energy and syllable duration.

Acknowledgements

We are most grateful for the contributions of Jill House, University College London, who made the recordings and transcriptions used in the experiments.

REFERENCES

[1] A. Cruttenden. Intonation, Cambridge University Press 1986.

[2] J. 't Hart, R. Collier, A. Cohen. A perceptual study of intonation, Cambridge University Press, 1990.

[3] D. Crystal. Prosodic systems and intonation in English, Cambridge University Press 1969.

[4] E. Couper-Kuhlen An introduction to English prosody, Edward Arnold 1986.

[5] A. Ljolje, F. Fallside. Recognition of isolated prosodic patterns using hidden Markov models, Computer Speech and Language 1987, vol 2, pp 27-33.

[6] J.W. Butzberger jr., M. Ostendorf, P.J Price, S. Shattuck-Hufnagel. Isolated word intonation recognition using hidden Markov models, Proceedings ICASSP 1990, pp. 773-776.

[7] G.J. Freij, F. Fallside. Lexical stress recognition using hidden Markov models, Proceedings ICASSP 1988, pp. 135-138.

[8] J.D. O'Connor, G.F. Arnold. Intonation of colloquial English, Longmans 1961.

[9] J.D. O'Connor, G.F. Arnold. Intonation of colloquial English, second edition, Longmans 1973.

[10] M.A.K. Halliday. A course in spoken English: intonation, Oxford University Press 1970.

[11] C. Jacobsen, B. Andersen. SIRtrain training software - user guide version 2.1, SUNSTAR - esprit project 2094, 1991.

[12] U. Jensen. Modelling of intonation contours - initial experiments & ideas, research note no. 186, Electronics division RSRE Malvern, January 1992.

[13] J.D. Markel, A.M. Gray. Linear prediction of speech, Springer-Verlag 1976.

[14] B.G. Secrest, G.R. Doddington. An integrated pitch tracking algorithm for speech systems, ICASSP 1983, pp. 1352-1355.

[15] waves+ (version 2) & DSP support, Entropic Research Laboratory, Inc. AT&T Bell laboratories 1991.

[16] J.R. Bellegarda, D. Nahamoo. Tied mixtures continuous parameter modeling for speech recognition, ASSP 1990, no. 12, pp. 2033-2045.

[17] U. Jensen. Modelling of intonation contours using continuous density hidden Markov modelling, master thesis, Aalborg University, August 1992.

[18] J. Kristiansen, B. Andersen, B. Lindberg. User guide to the SUNCAR recogniser V2.0, ESPRIT project 2094, SUNSTAR July 1992.

[18] S.J. Young. The use of syntax and multiple alternatives in the VODIS voice operated database inquiry system, Computer Speech and Language, vol. 5 1991, pp. 65-80.