MODELLING INTONATIONAL STRUCTURE USING HIDDEN MARKOV MODELS

Helen Wright and Paul Taylor

Centre for Speech Technology Research, University of Edinburgh,
80, South Bridge, Edinburgh, U.K. EH1 1HN
http://www.cstr.ed.ac.uk
email: {helen, pault}@cstr.ed.ac.uk

ABSTRACT

A method is introduced for using hidden Markov models (HMMs) to model intonational structure. HMMs are probabilistic and can capture the variability in structure which previous finite state network models lack. We show how intonational tunes can be modelled by separate HMMs and how HMMs can be used in a recognition system to automatically determine the tune type of an utterance.

1. INTRODUCTION

An intonational tune can be thought of as a distinctive contour type: utterances with the same tune sound the same intonationally, independent of the text of the utterance. Tunes have often been described in terms of idealised sequences of more atomic intonational units. For example, O'Connor and Arnold's [6] *high drop* consists of a low pre-head, high head and low fall, while Liberman's *surprise/redundancy* tune is L* H* L-L% [5].

Two approaches can be employed when producing a comprehensive inventory of tune types. Tunes can be classified in a bottom-up manner whereby similar patterns are grouped together and given abstract names relating only to their intonational properties. This is the basic approach of O'Connor and Arnold who describe ten canonical tunes, all with names such as *low bounce* and *high drop*. This paper examines the converse approach of top-down classification, whereby speech acts (or some other classification of utterance types) are assigned prototypical intonation patterns. This notion is often employed in speech synthesis systems where appropriate intonation must be assigned to yes/no questions, interjections and so on.

Top down approaches can be problematic because the variability associated with each type can make it difficult to specify contour types. While yes/no questions often rise at the end, we have to admit that a large number of examples don't, which makes a definitive statement on the behaviour of yes/no question intonation difficult.

Here we present a framework for intonational structure which explicitly takes into account the notion of variability, by using a likelihood model to describe tunes. Specifically, we use a hidden Markov model (HMM) to model each tune type. HMMs are probabilistic finite state networks (FSNs) and are the most commonly used technique in acoustic/phonetic modelling in automatic speech recognition and part-of-speech tagging.

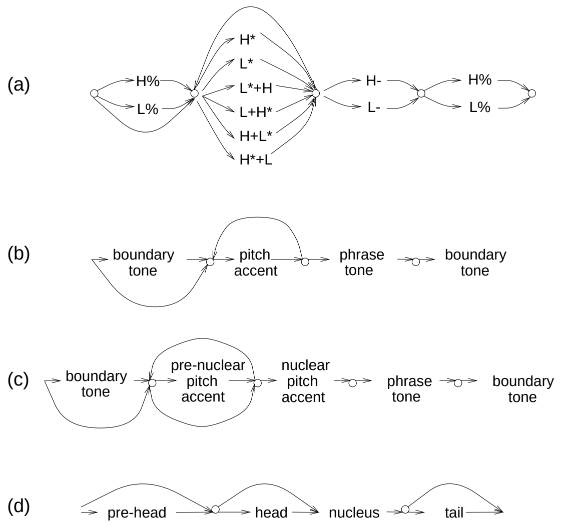

Figure 1. Intonational structure represented by finite state networks

Non-probabilistic finite state networks have previously been used to model intonational structure (and hence tune type) by specifying legal sequences of basic intonational elements, such as pitch accents or boundary tones. Figure 1a shows the familiar form of Pierrehumbert's intonational grammar giving all the legal tone sequences for English [7]. Figure 1b shows the same information but with discriptive variables associated with states which emit tones of the particular type (e.g. the pitch accent state emits all the pitch accent types). Figure 1c shows Ladd's [4] amended version where nuclear accents are treated differently from pre-nuclear accents. Figure 1d shows the British School system of pre-head, head, nucleus and tail. These FSNs represent deterministic structural descriptions of the tunes of English. A HMM is formed from these by adding probabilities. *Transition probabilities* are associated with arcs between states which give, for example, the likelihood of a contour having or not having a pre-head. *Observation probabilities* are associated with states and specify the likelihood of that state emitting one of the types associated with it. For example the pitch accent state in figure 1b might have a high chance of emitting a common accent such as H* and a much lower chance of emitting a rare accent such as H+L*. Each tune type is modelled by one HMM.

HMMs are both generators and acceptors of sequences of intonational elements. In generator mode they can be used to specify the distribution of elements for the HMM's tune type. In acceptor mode they can be used to give the probability that an observed sequence of intonational elements has been produced by that model. By comparing the probabilities of several models it is possible to determine which model was the most likely to have produced the observation sequence. The probabilities of an HMM can be automatically estimated by running the Baum-Welch training procedure on multiple training examples [2]. The remainder of our paper reports work which uses these aspects of the HMM for the task of assigning an appropriate tune type to an utterance whose type is not known.

2. RECOGNISING GAME MOVES IN THE MAPTASK

The experiments reported here use a subset of the DCIEM Maptask corpus [1] . This is a corpus of spontaneous goal-directed dialogue speech collected from Canadian speakers. The corpus has been analysed using the theory of conversational games first introduced by Power [9] and adapted for Maptask dialogues in Carletta et al. [3]. Conversational games are conventional sequences of acts, such as *question - answer - acknowledgement*, or, indeed, *request - non-linguistic-action - acknowledgement*. We distinguish 12 types of individual acts, which are termed "moves" in the conversational games.

The DCIEM corpus is fully spontaneous dialogue speech. In the experiments reported here, 20 dialogues (3726 moves) were used for training the system and 5 (1061) for testing. None of the test set speakers were in the training set, i.e. the system is speaker independent. The two participants in the dialogue have different roles called the *giver* and *follower*. Generally the giver is giving instructions and guiding the follower through the route on the map. Because of the different roles, each participant has a different distribution of moves.

Our hypothesis was that moves had characteristically different intonational tunes, and that HMMs could be used to determine the most likely move type for an unknown utterance. The ability to detect the most appropriate move for an utterance has proven very useful in the maptask automatic speech recogniser [11]. A key component of this system is the use of move specific language models. It has been shown that using a move specific language model on an utterance of the appropriate move type can reduce error rate. The HMM system described here is used to detect the move type for an unknown utterance, and based on this, a language model of that move type is used during recognition.

3. INTONATIONAL EVENTS AND TILT PARAMETERS

3.1. Problems with Discrete Intonational Classification

In sketching previous intonational systems in the HMM framework, we made the assumption that *discrete* HMMs were being used. That is, the observation probabilities for each state describe the probability of that state emitting one of the finite number of legal symbols for that state. In practice, we use *continuous density* HMMs in which states use a continuous probability density function to describe continuous variables rather than symbols.

Discrete intonational symbols have been avoided for a number of reasons. Firstly, we need a substantial quantity of hand-labelled data to train the HMMs. Even on clean speech human labellers find it notoriously difficult to categorise pitch accents reliably, and the reliability drops further for spontaneous speech. In a study [8] on the ToBI labelling (a variant of Pierrehumbert's scheme), labellers agreed on pitch accent presence or absence 80% of the time, while agreement on the category of the accent was just 64% and this figure was only achieved by first collapsing some of the main categories (e.g. H* with L+H*).

Secondly, we would need an inventory of labels which are suitable for describing intonational tune differences. In the ToBI scheme the distribution of pitch accent types is often extremely uneven. In a portion of the Boston Radio news corpus which has been labelled with ToBI, 79% of the accents are of type H*, 15% are L*+H and other classes are spread over the remaining 6%. From our point of view such a classification isn't very useful because some tunes which are clearly different will be marked with H* throughout. Furthermore, intonational factors which are significant in tune description, such as local prominence, are omitted.

Thirdly, we would need an automatic system capable of producing the symbol sequence for an utterance. Similar to human labellers, automatic systems (e.g. [10]) find the task of locating pitch accents much easier than classifying them.

3.2. Intonational Events and Tilt Parameters

Rather than have several discrete symbols to describe intonation, we have one for accents, one for boundaries and a combined one for when accents and boundaries occur too close to be separated. These are termed *intonational events* and carry linguistically interesting intonational information. In addition we use a label for silence and a label "connection" to represent any part of the contour that is not classified as one of the events. A diacritic of "minor" was used to mark accents which were either small or whose existence was questionable. The 25 dialogues used for training and testing were hand labelled using this scheme.

To describe the intonational content of the pitch accents and boundaries we use 4 continuous variables collectively known as *tilt* parameters [12]. These are *start F0*, which is the F0 value at the start of the event; *amplitude*, a measure of the F0 excursion of the event; *duration* (in time); and *tilt*, a continuous dimensionless parameter expressing the shape of the event (a value of -1 means the event is a pure fall, +1 means a pure rise and values between indicate the event has a rise and fall). These values can be calculated automatically given the approximate location of a event (accent or boundary) and the F0 contour.

4. USING HMMS TO RECOGNISE MOVE TYPE

In the speech recognition application described above, the process of recognising moves from the data must be fully automatic and hence we must derive the intonational events themselves automatically. Intonational event detection is performed by using a continuous density HMM system (which is completely separate from the move type detector).

4.1. Locating and Analysing Intonational Events

Each utterance is represented acoustically by F0 and energy, and their first and second derivatives. A single context independent model is trained for each of the main label categories. The system is trained on the hand labelled data described in section 2.

Performance is assessed by measuring how well the hand labelled test set matches the output of the recogniser. Only accents and boundaries are counted. Silence is unimportant and connections are positioned as a consequence of accent and boundary placement and hence are redundant. For an automatically labelled event to count as correct, it must overlap a hand labelled event by at least 50%. Using this metric the performance of the recogniser is 74.3% correct with an accuracy of 29.4%. A large number of errors arose from minor accents. When these were ignored the performance was 86.5% correct with 54.3% accuracy.

The low general accuracy is almost certainly a result of the data being spontaneous and speaker independent. An equivalent speaker dependent system trained on part of the data gave 87% correct and 63% accuracy, while a system trained on fluent "simulated dialogue" speech gave 85% correct with 76% accuracy. We are currently examining speaker normalisation techniques which we hope will increase performance on the speaker independent data.

After this stage, each event is analysed to determine its tilt parameters.

4.2. Experiments

We use a different HMM to model the intonation of each type of move. As observations, the HMMs use sequences of vectors, each corresponding to the 4 tilt parameters of a single event. A three state, left-right continuous density HMM is trained for each move. The observation density functions comprise of a set of weighted gaussians, each with a mean and variance.

Model	% correct
Unigram	39.5
Bigram	44
Giver Bigram	49
Follower Bigram	55
Giver+follower Bigram	61

Table 1. Results for 12 game moves

The HMMs are trained in two stages, *initialisation* and *re-estimation*. Initialisation involves providing crude estimates for the HMM parameters, which are then iteratively re-estimated using the standard Baum-Welch algorithm. Two types of initialisation (described in section 5.) were investigated.

Preliminary experiments showed that moves follow one another with some degree of predictability (e.g. a reply-yes or reply-no is the most common response to a query-yes/no). We make use of this by combining the HMMs with an N-gram model which gives the a priori probability of a sequence of N moves occurring.

The data was modified in a number of ways, omitting either silences or connections or both. The modified data improves the recognition results as it omits linguistically non-significant events. The data was also normalised in an attempt to eliminate some speaker specific characteristics, such as overall pitch range. All the observation vectors for a particular speaker were normalised by subtracting the mean and dividing by the standard deviation for that speaker.

4.3. Results

Table 1 gives a summary of results for recognition on the 12 game moves. If utterances are considered in isolation (achieved by using a unigram), 39.5% of moves are recognised correctly. The type of the previous move of the other speaker can be used to condition the choice of the move under examination by using a bigram. Using this the score increases to 44% correct.

This analysis corresponds to an overhearer scenario where one is trying to determine the move type of both participants in the dialogue. The type of the previous move has been guessed automatically and hence the probability of the current move may be conditioned on incorrect information. An alternative scenario is where the move type of one participant is known, for example in human-computer dialogue. In this case (imagining that we are recognising the speech of the human participant) the move type of the computer's previous utterance is known, and hence we have a better chance of a priori estimation of the current utterance. The distribution of move types is different for givers and followers and hence we can use different bigrams depending on whether we are recognising the giver's or follower's moves. Under these conditions the performance improves to 49% when recognising the givers' moves and 61% when recognising the followers' (54% overall).

5. DISCUSSION

In order to examine exactly how the HMMs model the intonation contour, we investigated the relationship between states and different types of intonational events. Two types of HMM initialisation were examined. In the first type, the sequence of observations is divided into three equal sized parts and each state estimates its parameters on one of these parts. The second type gives each state an explicit function: state 1 models prenuclear events (head), state 2 nuclear accents and state 3 boundary events (tail), reflecting the intonation contour structure of the British School and Ladd. Recognition scores of HMMs after initialisation are poor, but improve greatly after re-estimation. Re-estimation is an unsupervised iterative technique which optimises the maximum likelihood of the models emitting the observations in the training data. The re-estimation process takes several iterations and often after training the states do not emit the same observations as after initialisation.

The overall likelihoods of each state emitting different types of events when recognising each of the test utterances were investigated. We found that both types of initialisation produced similar recognition results with similar state occupation statistics. State 1 mostly emits pre-nuclear accents, states 2 and 3 both emit nuclear accents and state 3 mostly emits boundary tones. A possible reason why state 3 emits nuclear accents is due to the fact that many of these are combined accents and boundary tones. While there is no exact alignment of states to event types, this shows that the HMMs are modelling the contour in more or less the same way as the FSNs of the British School and Ladd. Although the two types of initialisation process are different, the re-estimation procedure produces similar HMMs regardless.

The results reported above are for a fully speaker independent task. A previous pilot study which had the same speakers in the test and training sets produced a unigram score of 49% and a bigram score of 55% for the basic task, compared with 39% and 44% shown above.

In order to investigate possible intonational similarities between move types, we clustered the moves. One experiment grouped the 12 moves into 3 new types, termed *questions*, *statements* and *replies*. A unigram model gave 68% correct and a bigram model 70%.

6. CONCLUSION

The main purpose of this work is to show that HMMs are a suitable model for intonational structure, and that this facilitates practical applications such as recognising tune types from F0 contours. The specific task described here was motivated by the fact that the move recogniser is used in our speech recognition system. We hypothesised that each move has a distinct tune but the fact that the recognition scores are not perfect shows that this is a weak assumption. Better scores should be obtainable if the utterance types being recognised correlate more directly with intonational behaviour, as was the case when the moves were clustered. It should be clear that many other types of utterance classification (both top down and bottom up) are compatible with the HMM framework. Recognition results are dependent on the number of tune types being modelled and whether the tune classification is top-down or bottom-up.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support from the UK Engineering and Physical Science Research Council through EPSRC grant GR/J55106 and ESPRC PhD studentship 9630715.

REFERENCES

[1] E. G. Bard, C. Sotillo, A. H. Anderson, and M. M. Taylor. The DCIEM map task corpus: Spontaneous dialogues under sleep deprivation and drug treatment. In *Proc. of the ESCA-NATO Tutorial and Workshop on Speech under Stress, Lisbon*, 1995.

[2] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process. *Inequalities*, 3:1–8, 1972.

[3] J. Carletta, A. Isard, S. Isard, J. Kowtko, G Doherty-Sneddon, and A H. Anderson. The coding of dialogue structure in a corpus. In J.A. Andernach, S.P. van de Burgt, and G.F. van der Hoeven, editors, *Proceedings of the Ninth Twente Workshop on Language Technology: Corpus-based Approaches to Dialogue Modelling*. Universiteit Twente, Enschede, 1995.

[4] D. R. Ladd. *Intonational Phonology*. Cambridge Studies in Linguistics. Cambridge University Press, 1996.

[5] M. Y. Liberman. *The Intonational System of English*. PhD thesis, MIT, 1975. Published by Indiana University Linguistics Club.

[6] J. D. O'Connor and G. F. Arnold. *Intonation of Colloquial English*. Longman, 2 edition, 1973.

[7] J. B. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT, 1980. Published by Indiana University Linguistics Club.

[8] J. F. Pitrelli, M. E. Beckman, and J. Hirschberg. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *ICSLP94*, volume 1, pages 123–126, 1994.

[9] R. Power. The organization of purposeful dialogues. *Linguistics*, 17:107–152, 1979.

[10] K. Ross and M. Ostendorf. A dynamical system model for recognising intonation patterns. In *EUROSPEECH 95*, 1995.

[11] P. A. Taylor, S. King, S. D. Isard, H. Wright, and J. Kowtko. Using intonation to constrain language models in speech recognition. In *Eurospeech 97*, 1997.

[12] P. A. Taylor and A. W. Black. Synthesizing conversational intonation from a linguistically rich input. In *Second ESCA/IEEE Workshop on Speech Synthesis, New York, U.S.A*, 1994.