USING INTONATION TO CONSTRAIN LANGUAGE MODELS IN SPEECH RECOGNITION

Paul Taylor, Simon King, Stephen Isard, Helen Wright and Jacqueline Kowtko

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

ABSTRACT

This paper describes a method for using intonation to reduce word error rate in a speech recognition system designed to recognise spontaneous dialogue speech. We use a form of dialogue analysis based on the theory of conversational games. Different move types under this analysis conform to different language models. Different move types are also characterised by different intonational tunes. Our overall recognition strategy is first to predict from intonation the type of game move that a test utterance represents, and then to use a bigram language model for that type of move during recognition.

1 INTRODUCTION

This paper describes a method for using intonation to reduce word error rate in a speech recognition system designed to recognise spontaneous dialogue speech. Our experiments are on the DCIEM Maptask corpus [2], a corpus of spontaneous task-oriented dialogue speech.

Our dialogue analysis is based on 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. We show in section 4 that if a separate bigram language model is trained for each move type, the bigram entropy for most types is lower than the entropy for a general bigram model. Thus move-specific grammars are more constrained than general language models. Our overall recognition strategy is first to predict the type of game move that a test utterance represents, and then to use a bigram grammar for that type of move during recognition.

This work had its origin in the intuition, borne out in [6], that different types of move are characterised by different intonational tunes. Not that there is a perfect one-to-one correspondence between move types and tunes, but rather that intonation should help to distinguish among the various move types that can occur at a given point in a game.

For instance, a *checking* move like "Now, you've got an old barn on your map, don't you?", that asks the addressee to confirm something that the speaker is pretty sure is true, expects an affirmative answer without much intonational marking. A negative reply or a request for clarification would have more prominent pitch markings.

So we expect the *combination* of intonation and dialogue context to be a good predictor of move type. However, in this paper we investigate the extent to which intonation alone can act as a useful predictor of move type, and hence as a constraint for language modelling.

2 DATA

The experiments here use a subset of the DCIEM Maptask corpus. This is a corpus of spontaneous goal-directed dialogue speech collected from Canadian speakers. The data files were coded using the conversational game analysis and split into smaller more manageable files, with one move per file. The files were also hand labelled with the intonation scheme described in section 3.1. 20 dialogues (3726 moves) were used for training, and 5 dialogues (1061 moves) were used for testing. None of the test set speakers were in the training set. The intonation recogniser, the move recogniser, the language models and the HMM phone models were all trained from the same training data.

3 MOVE TYPE DETECTION FROM INTONATION

3.1 Intonational Event Recognition

Intonation is characterised using 4 basic intonational labels: **a** for pitch accents, **b** for boundaries, **c** for connections and **sil** for silence. The **a** and **b** labels are called intonational *events* and represent the linguistically significant portion of the intonation contour. **c** is used simply to fill in parts of the contour which are not an event or silence. In addition, a compound label **ab** is used when an accent and boundary are so close they overlap and form a single intonational event.

In this system, unlike others such as ToBI [12], there are no distinct categories of pitch accents and boundaries. Discrete intonational categories have been avoided for a number of reasons. Firstly, even on clean speech, human labellers find it notoriously difficult to label the categories reliably, and the reliability drops further for spontaneous speech. In a study on ToBI labelling [8], 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, 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 an information theoretical point of view, such a classification isn't very useful because virtually everything belongs to one class, and therefore very little information is given by accent identity. Thirdly, recognition systems which have attempted to automatically label intonation usually do much better at the accent detection task than at classifying the accents. Ross and Ostendorf [10] describe a system which is very successful at pitch accent detection (85%-89%), but from examination of the confusion matrix, is it clear that the system is really only detecting high (the family of H*) accents accurately - the fact that other types are difficult to detect and are often confused with high accents doesn't affect the score too much because of the overwhelming proportion of high accents in their data.

Thus we choose a single category of accent, because both human and automatic labellers find it difficult to distinguish more, and because the amount of information obtained even if accuracy were high would be small. To put it another way, in practical situations the ToBI system more or less equates to a single pitch accent type anyway - all we have done is to make this explicit.

However, this is not to say that we believe that all pitch accents are identical, rather that current categorical classification systems aren't suited for our purposes. To classify pitch accents, we use 4 continuous variables collectively known as *tilt* parameters [14]. 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 give the approximate location of a event (accent or boundary) and the F0 contour.

Our automatic event detector is based on a continuous density HMM system. 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 hand labelled data.

We assess performance by measuring how well the hand labelled test set matches the output of the recogniser. Only accents and boundaries are counted as silence is unimportant and connections are where they are 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%. The low 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 will hopefully increase performance on the speaker independent data.

3.2 Move Detection

It is generally accepted that intonation contours do not consist merely of a string of intonation events, but have an internal structure. For example, the British School [4] divides the contour into head, nucleus and tail and Pierrehumbert and Ladd have also proposed finite state grammars specifying possible sequences of intonational events [7]. Rather than using deterministic finite state networks, we adopt a stochastic approach by adding probabilities to the networks, and use hidden Markov models. Each state of these HMMs is intended to capture the characteristics of a different part of the intonation contour (e.g. head, nucleus).

We have postulated that different types of move will have different intonational characteristics and use a different HMM to model the intonation of each type of move. As observations, the HMMs use the output from the intonation event recogniser, which consists of a sequence of vectors of tilt parameters. Each vector represents a single intonational event. A three state, left-right continuous density HMM is trained for each move.

The move type recogniser is combined with an N-gram model, which gives the a priori probability of a sequence of moves occurring. To date we have only implemented a unigram model in the full system, but preliminary experiments have shown that move bigrams improve performance by making use of the fact 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).

The baseline system recognises 34% of moves correctly. By adjusting the grammar scaling factors of the unigram this increases to 38%. Again we believe that speaker variability is a major factor in producing errors: in a speaker dependent study performance was around 55%. We achieved some increase in performance by simply normalising the tilt feature vectors for each speaker, but it this does not remove all speaker specific effects.

Given the original theoretical motivation for using HMMs to model intonational tunes, we conducted additional experiments involving the initialisation in training of particular events to specific states, namely pre-nuclear events (head) to state 1, nuclear accents to state 2 and boundary events (tail) to state 3. This reflects the British School's traditional intonation contour structure. The results remained the same at 38% correct. Informal inspection of the HMM states indicates that state 2 of the models does assign higher probability to large pitch excursions than the other states, and so it is to some extent modelling nuclear accents.

4 MOVE TYPE SPECIFIC LANGUAGE MODELLING

To test our hypothesis that each move type has a distinct language model, we estimated backed off bigram [5] language models for each of the 12 types. The training/testing sets for the language models were the same as those used for other parts of the system. The number of training tokens per move type varied from 300 to 8000 and the vocabulary size is ∼900. A general backed off bigram was also estimated on all training data (∼24 000 tokens).

The sentences in the test set were grouped into 12 subsets depending on their move type. The perplexity of the general bigram was calculated and compared with the perplexity of the appropriate move specific bigram on each subset. The perplexities of the various language models on the test set are given in Table 1. The perplexity based on always using the appropriate move-specific language model appears in the table as "100% detection".

In many cases the perplexity of the move specific bigrams was significantly lower than that of the general bigram. The overall perplexity when using move-specific grammars is slightly lower than when using the general grammar. By using the general model for the move types where it did better, the overall perplexity can be further reduced – this is shown in the table as "best choice".

The amount of data available to estimate the move type specific language models (mean ∼2 000 tokens per move type) is not really adequate. To gauge the effect of this data shortage, we generated a language model on just $^1/_{12}$ of the training data (∼2 000 tokens). This grammar has much higher perplexity than any other language model.

	grammar	perplexity
general grammar	all data	27.66
	$^1/_{12}$ data	43.14
move specific	100% detection	27.08
	best choice	25.55

Table 1. Perplexities of the various language models

5 RECOGNITION EXPERIMENTS

5.1 Recogniser

A standard HMM-based speaker-independent cross-word 8-mixture triphone speech recognition system was used in all experiments. We use several special "words" for non-speech and word sequences. Whilst these are treated like any other word for language modelling and recognition purposes, they are ignored in the calculation of recognition accuracy.

5.2 Experiments

The general-purpose language model was used as a baseline, and achieved 61.3% word accuracy.

We then used move type specific language models. In the first experiment, the correct move type was used for each test utterance (100% move detection), and in a second experiment the automatic move type recogniser (see section 3.2) was used. In this second experiment, only the most likely move type for each utterance was considered. This thresholding is not ideal, and is discussed in section 6. For comparison, we also used the general purpose language model estimated from $^1/_{12}$ of the data. In a final experiment, we assigned a random[1] move type to each utterance. The results are shown in table 2.

	grammar	% Word Accuracy
general grammar	all data	**61.3**
	$^1/_{12}$ data	58.3
move specific	100% detection	62.7
	best choice	63.1
	automatic	**58.5**
	random	52.5

Table 2. Recognition results

6 DISCUSSION

Our hypothesis was that intonation is a good predictor of move type and that move specific language models are a useful constraint for speech recognition.

6.1 Intonation

We can infer from the results in table 2 that intonation *does* carry useful information, because the automatic move detection system gave improved results (58.5%) over the random move type experiment (52.5%) and over the grammar trained on $^1/_{12}$ of the data (58.3%). The poor performance of the system which assigned random move types to utterances shows both that the intonation recogniser is adding useful information, and that the language models for the various move types are distinct. However, the automatic system did not do as well as that with the grammar trained on all data. We believe this is mainly due to insufficient data in move specific language model training.

The HMM approach to intonational tunes outperforms other approaches we have tried, for instance the neural net system described in [13]. We have represented intonation with continuous tilt parameters in experiments described here and hence used continuous density HMMs. It is worth noting that other intonational representations can also be modelled by HMMs and specifically a representation based on discrete categories of accents could

[1] keeping the distribution of the types the same as in the 100% move detection case

be modelled by a HMM with discrete observation probabilities.

6.2 Language modelling

It is clear that the move types do form clusters in terms of language model. This can be seen from table 2; the 100% move type detection result (62.7%) is better than that for the general purpose grammar (61.3%). Further language model improvements we intend to examine include smoothing the grammars using a grammar trained on all data, or on other corpora, such as the HCRC Map task [1]. Although the 12 move types used here did exhibit some degree of clustering, we feel that they are probably not the optimal choice. Merging or splitting of some move types may give improved results.

6.3 Further Work

As mentioned in 5.2, the results we quote are for a system where a single language model is chosen on the basis of the intonational analysis, and speech recognition is governed by that language model. We expect better results from an integrated system which computes the probability of recognition hypotheses according to all models, weighted according to the intonational probability of each model for the utterance under consideration.

We will also consider context-dependent intonational models, as discussed in the Introduction, and higher order N-grams for a priori move probabilities. This can be done in either in "participant" mode, where for each utterance the system models the other speaker, and hence knows for certain what type of move has just been made previously, or "overhearer" mode, where it performs a Viterbi search for the best move sequence over the entire conversation. Participant mode can of course be expected to give better results, consistent with studies such as [11].

7 ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the UK Engineering and Physical Science Research Council through EPSRC grant GR/J55106.

REFERENCES

[1] Anne H. Anderson, Miles Bader, Ellen G. Bard, Elizabeth H. Boyle, Gwyneth M. Doherty, Simon C. Garrod, Stephen D. Isard, Jacqueline C. Kowtko, Jan M. McAllister, Jim Miller, Catherine F. Sotillo, Henry S. Thompson, and Regina Weinert. The hcrc map task corpus. *Language and Speech*, 34(4):351–366, 1991.

[2] Ellen G. Bard, Catherine Sotillo, Anne 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.

[3] Jean Carletta, Amy Isard, Stephen Isard, Jacqueline Kowtko, Gwyneth Doherty-Sneddon, and Anne 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] David Crystal. *Prosodic Systems and Intonation in English*. Cambridge Studies in Linguistics. Cambridge Universtity Press, 1969.

[5] Slava M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recogniser. *IEEE Trans. Acoustics, Speech and Signal Processing*, 35(3):400–401, 1987.

[6] Jacqueline C. Kowtko. *The Function of Intonation in Task Oriented Dialogue*. PhD thesis, University of Edinburgh, 1996.

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

[8] John F. Pitrelli, Mary E. Beckman, and Julia 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] Ken Ross and Mari Ostendorf. A dynamical system model for recognising intonation patterns. In *EUROSPEECH 95*, 1995.

[11] M. F. Schober and H. H. Clark. Understanding by addressees and overhearers. *Cognitive Psychology.*, 21:211–232, 1989.

[12] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labelling English prosody. In *Proceedings of ICSLP92*, volume 2, pages 867–870, 1992.

[13] P. A. Taylor, H. Shimodaira, S. D. Isard, S. King, and J. Kowtko. Using prosodic information to constrain language modesl for spoken dialogue. In *ICSLP'96, Philadelphia*, 1996.

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