Automatic Utterance Type Detection Using Suprasegmental Features

Helen Wright

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

ABSTRACT

The goal of the work presented here is to automatically predict the type of an utterance in spoken dialogue by using automatically extracted suprasegmental information. For this task, we present and compare three stochastic algorithms: hidden Markov models, artificial neural nets, and classification and regression trees. These models are easily trainable, reasonably robust and fit into the probabilistic framework required for speech recognition. Utterance type detection is dependent on the assumption that different types of utterances have different suprasegmental characteristics. The categorisation of utterance types is based on the theory of conversation games and consists of 12 move types (e.g. reply to a question, wh-question, acknowledgement). This utterance type detector is used in an automatic speech recognition system to reduce the word error rate.

1. INTRODUCTION

This paper describes a method of automatically detecting the type of an utterance, using only prosodic information, for use in an automatic speech recognition system. Suprasegmental features are automatically extracted and used to train a stochastic model for each of 12 utterance types. However, there is not always a one-to-one mapping of intonation and utterance type, for example, a yes/no question frequently has a high boundary tone, but sometimes has a low one. As long as each utterance type has different distributions of observed suprasegmental features in the training data, the stochastic models for each move type will differ. When confronted with a set of prosodic features of a previously unseen utterance, we can calculate the likelihood that each of the models produced it. These likelihoods are then used to determine the probability of each utterance type, given its suprasegmental features.

This paper investigates and compares the effectiveness of three different stochastic models for the task described above, namely hidden Markov models (HMMs), classification and regression trees (CART) and artificial neural nets (ANN).

The utterance type detector is used in an automatic speech recognition system to select a specific language model thus reducing word error rate. As well as applications in automatic speech recognition systems, an utterance type detector can be used in human-computer dialogue systems, for example to determine whether the conversation agent is being asked a question or not. Other applications include automatic summarisation and machine translation of conversations.

2. DATA

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. This Maptask corpus was chosen as it is readily available, easy to analyse, has a limited vocabulary and structured speaker roles.

The DCIEM corpus is fully spontaneous dialogue speech. 20 dialogues (3726 moves) are used for training the system and 5 (1061) for testing. None of the test set speakers are 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. Due to the different nature of the roles, each participant has a different distribution of moves.

The corpus has been analysed using the theory of conversational games first introduced by Power [6] and adapted for Maptask dialogues in Carletta et al. [3]. The Games Analysis for the Maptask corpus consists of six games: *Instructing, Checking, Query-YN, Query-W, Explaining* and *Aligning*. The games consist of initiating moves (*instruct, check, query-yn, query-w, explain and align*) and non-initiating moves (*acknowledge, clarify, reply-yes, reply-no, reply-w, ready*).

3. SYSTEM ARCHITECTURE

This paper describes one module in the automatic speech recognition system described in [9]. In order to reduce word recognition error, one can take advantage of certain regularities in syntax and lexical distributions associated with different utterance types. For example, a yes/no question frequently starts with “Do you have ...”. Language models are used in Automatic Speech Recognition systems to constrain the number of word recognition possibilities. Move specific language models are more effective at this task as they have a smaller variation of possible word sequences and hence lower perplexity (c.f.[9]).

The appropriate language model is chosen by calculating the most likely move type (M) given the suprasegmental features of the utterance (I), i.e. the utterance with the highest posterior probability, $P(M|I)$. To calculate this, one has to calculate the prior probability of the move $P(M)$ and multiply it by the output of the likelihood model $P(I|M)$, formalised in the Bayesian formula:

$$P(M|I) = P(M)P(I|M)$$

A *dialogue model* is trained to predict the prior probability, $P(M)$, of sequences of moves. For instance, a query followed

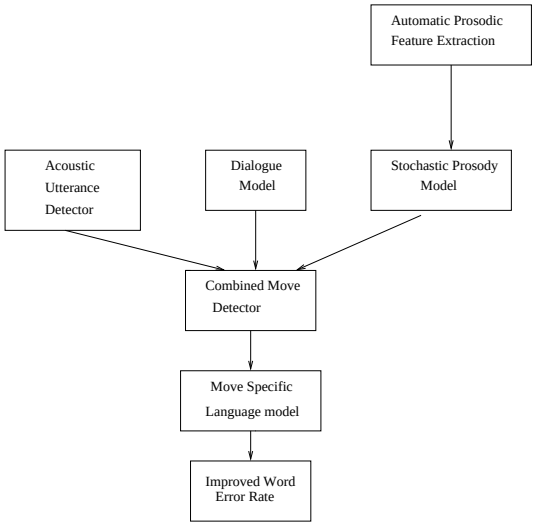

Figure 1: System architecture of an automatic speech recognition system using utterance type specific language models

by a response, followed by an acknowledgement is more likely than three acknowledgements in succession. A unigram is the simplest type of dialogue model, this reflects the likelihood of a move given its distribution in the training dialogues. The dialogue model that has the best predictive power (i.e. reduces the perplexity of the test set the most) is a 4-gram model. This uses the identity of the current speaker and the speaker of the previous move, and the last move of the previous speaker.

In separate experiments, three types of likelihood models (HMMs, CART and ANNs) are used to model the different prosodic characteristics of the move types. These are used to calculate the likelihood of a set of observed suprasegmental features for each move type $P(I|M)$. Each of these three models are discussed and their effectiveness for the given task compared.

As described in [9] and [4], the model trained on suprasegmental features is used in conjunction with a move detector trained on acoustics. The recogniser is run once and the output word sequence is used to estimate the likelihood of a move. A viterbi search finds the most likely path through the dialogue model, given the observations from the suprasegmental and acoustic models. The probability of a sequence of moves is the product of the *transition probability* (given by the dialogue model) and the *state probability* which is the weighted sum of the prosodic and the acoustic models. By varying these weights one can place more emphasis on one or the other models. The remainder of this paper reports the best method for maximising the accuracy of the likelihood estimation model trained on prosodic features, thus improving move detection accuracy and therefore reducing word recognition error.

4. INTONATION EVENTS AND TILT PARAMETERS

All three stochastic models (HMM, CART and ANNs) use features extracted from *intonation events*, which are automatically recognised by an algorithm described in [8]. Intonation events are categorised as: a (pitch accent), b (boundary tone), and ab (for when an accent and boundary co-occur). The system is trained on the data described in section 2, hand-labelled for intonation events. A single context independent hidden Markov model is trained for each of the event types (a, b and ab), using as observations F0 and rms energy at 10ms intervals, together with standard rate of change and acceleration measures ("deltas"). The means and variances for each speaker's F0 and energy are calculated and used to normalise the data for that speaker. Once trained, the system is run by using the HMMs for each label in combination with a bigram model representing the prior probabilities of pairs of labels occurring in sequence. The viterbi decoding algorithm is used to determine the most likely sequence of labels from the acoustics of the utterance being recognised. This system identifies 86.5% of the hand-labelled events correctly.

To capture the characteristics of the intonation events, each one is parameterised in terms of 4 continuous variables, known as *tilt parameters* [8]. These are *start F0* (in Hertz), which is the F0 value at the start of the event; *amplitude* of the F0 excursion of the event (Hertz); *duration* (seconds); and *tilt*. Tilt is a continuous dimensionless parameter expressing the shape of the event. The tilt parameter has a range of -1 to 1, where -1 is pure fall, 1 is pure rise, and 0 contains equal portions of rise and fall. The values of the 4 parameters are calculated automatically given the approximate location of an event and the F0 contour. The mean and standard deviation of each tilt parameter (excluding tilt) is calculated for each speaker. These are used to normalise the parameters for each event in order to reduce speaker variation effects.

5. STOCHASTIC MODELLING OF SUPRASEGMENTAL FEATURES

5.1. Hidden Markov Models

Each move has a sequence of intonation events. We model these sequences by using a separate state for the beginning, middle and end of utterances. We use a *hidden Markov model* because the state sequence is not deterministically recoverable from the observation sequence. By using a viterbi decoder at run time, the most probable state sequence is determined, given the observation sequence. The HTK, hidden Markov toolkit is used to train and test the HMMs [11]

The parameterised events form 4 component observation vectors for continuous density hidden Markov models, with one model for each of the 12 move types. The model consists of 3 states with transitional arcs as illustrated in figure 2. *Observation probabilities* ($b_j(o_t)$) specify the likelihood of a state emitting an event, whose tilt values are described by the continuous density function associated with that state. The observation density function is a 6-component Gaussian mixture. *Transitional probabilities* (a_{ij}) are associated with the arcs between each state and determine the state transitions, depending on the position in the utterance.

HTK [11] allows observation vectors to be split up into a number of independent data streams. These streams can be weighted, enabling one to posit more importance on one or two of the parameters. The best results are obtained by combining start F0 and

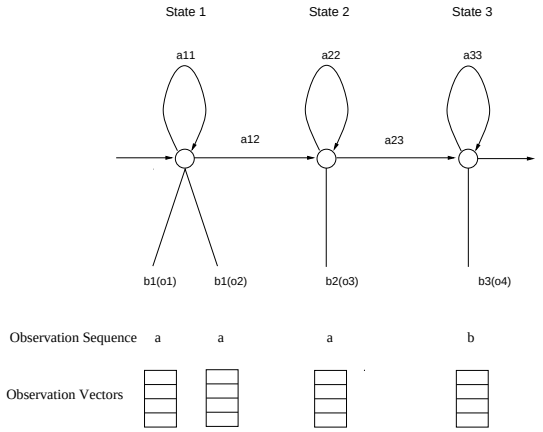

Figure 2: A three state, left-to-right HMM

F0 amplitude as one stream and giving it slightly more weighting than duration and tilt which are combined as a second stream.

The hidden Markov models are trained using the Baum-Welch algorithm to provide transition and observation probabilities for modelling their particular training data. Once trained, the HMMs are run over each utterance and the HMM that matches the utterance the closest is chosen as the answer.

5.2. Classification and Regression Trees

This section describes work inspired by the use of classification and regression trees trained for utterance type detection on a large database of telephone conversations, reported in [7]. They use similar features to train a classification tree to differentiate 5 utterance types with reasonable success.

54 suprasegmental and durational features are used to construct tree structured classification rules, using the CART training algorithm [2]. The trees can be examined to determine which features are the most discriminatory in move classification. The output of the classification tree is the probability of the move given the features, i.e. the posterior probability $P(M|I)$. In order to compare the trees with the HMMs, the likelihood of observing a set of features given a certain move $P(I|M)$, is calculated by dividing the output of the tree by the output of the unigram, i.e. the prior probability $P(M)$. An alternative method is to train the tree on data containing equal numbers of moves. Both methods produce similar results.

The suprasegmental features are automatically extracted from the speech signal and used to train the classification tree. For each move the last three accents (if present) are automatically detected and their 4 tilt parameters extracted and normalised, as described in section 4. The other prosodic features are based on F0 (e.g. max F0, F0 mean and standard deviation), rms energy (e.g. energy mean and standard deviation) and duration (e.g. number of frames in utterance, number of frames of F0). These features capture general characteristics of the utterance, for example the standard deviation of the F0 represents pitch range.

As the final part of the intonation contour is often indicative of utterance type, similar calculations are made for the last and penultimate 200ms of the utterance (e.g. mean RMS energy in the end region normalised using the mean and standard deviation of RMS energy for the whole utterance). Other features are calculated by comparing feature values for the two end regions and the whole utterance (e.g. ratio of mean F0 in the end and penultimate regions, difference between mean RMS energy in the end and penultimate regions). In addition to these features the least-squares regression line of the F0 contour is calculated for the last 200ms and for the whole utterance. This would capture intonation features such as declination over the whole utterance, and boundary type over the final part of the contour.

Feature Type	Usage (%)
Duration	0.47
F0	0.41
RMS Energy	0.12

Table 1: Discriminatory features and type usage in move classification

It is useful to know which features are the most discriminatory in the classification of the moves. As the tree is reasonably large with 30 leaves, interpretation is not straightforward. For simplicity, we group the features into 3 general categories of duration, F0 and energy. Table 1 gives the *feature usage frequency* for these groups of features. This measure is the number of times a feature is used in the classification of data points. It reflects the position in the classification tree as the higher the feature is in the tree, the more times it will be queried. The measure is normalised to sum to 1 for each tree.

Different moves types by their nature vary in length, therefore it is not surprising that duration is highly discriminatory in classifying utterance types. For example, ready, acknowledge, reply-yes, reply-n and align are distinguished from the other moves by the top node which queries a duration feature. This duration feature, regr_num_frames, is the number of frames used to compute the F0 regression line for a smoothed F0 contour over the whole utterance. This is comparable to the study reported in [7], where durational features were used 55% of the time and the most queried feature was also regr_num_frames. This feature may be a fairer measure of actual speech duration as it excludes pauses and silences.

The F0 features that come highest up in the tree are F0 mean in the end region, maximum F0 and tilt value of the last accent. This indicates that the F0 near the end of the utterance contains important linguistic information for the distinction of utterance types.

5.3. Neural Nets

The 54 features described in section 5.2 are used as input for a three layer perceptron neural network. The input layer consists of 54 nodes, one for each of the features which are normalised to fall between 1 and -1. The network contains one hidden layer of 60 units. The output layer consists of 12 nodes, one for each of the moves. Whichever node has the highest activation value is taken as the most likely move type. The net is trained with stochastic back propagation algorithms using a cross entropy cost function.

	HMM	CART	ANN
unigram on all moves	42	44	43
unigram on initiating moves	36	39	36
unigram on other moves	48	49	50
4-gram on all moves	64	63	62
4-gram on initiating moves	56	55	54
4-gram on other moves	72	71	70

Table 2: Percentage of moves correctly recognised

The system used to train and test the neural nets is the Stuttgart Neural Network Simulator [10].

ANNs, like classification trees, factor in the distribution of moves in the training database, estimating the posterior probability. In order to compare the ANNs with the other systems, we derive the likelihood $P(I|M)$, by dividing the output of the ANN by the prior probability.

6. RESULTS

Table 2 gives the results for the number of moves correctly recognised by the different stochastic models, when combined with a unigram dialogue model. All three methods obtain similar move recognition results, with the CART method doing slightly better with 44% correct. If one was to always choose the most frequent move *acknowledge*, one would get 25% of the moves correct. The stochastic models obtain results significantly above this figure. All three models do better on non-initiating moves than initiating moves. This is useful in human-computer interaction systems where the the word recognition accuracy is not as important as knowing the type of response. For example, differentiating "yeah, yes, yep" is not as important as the fact that the utterance is a positive reply to a question.

By combining the intonation model with the 4-gram described in section 3, the accuracy is increased to 64% for the HMMs, 63% for CART and 62% for the Neural Nets. The baseline result for the system is 24.8% word error rate, obtained by using a general language model. By using only the acoustic model for move recognition, the best WER obtained is 24.1% with 40% move detection accuracy with a unigram, increasing to 57% with the 4-gram but with no improvement of WER. The ANNs do not show an improvement of the baseline WER, obtaining 26.15%. This may be due to the uneven distribution of move types in the training data, resulting in some poorly trained classes. HMMs obtain 23.7% word error rate and CART 23.6%. This reduction of the baseline system word error rate is significant.

7. FURTHER WORK

Possible ways of further reducing the word error rate include having a better utterance type set, more sophisticated language models and as with all speech recognition systems, more data.

One way to improve the utterance type set would be with context specific intonation models. For example, one can postulate that the intonation patterns may vary between a *negative reply* to a *yes/no question* and one to a *check* question that expects a positive reply. Kowtko [5] shows different intonation patterns for one move type (acknowledge) in different contexts. Training context specific models, however, are problematic due to the sparseness of the data.

One could divide the dialogue into a set of utterance types, whose corresponding language models improve the speech recognition the most. There is no guarantee, however, that this set will be meaningful in dialogue terms or be distinguishable in terms of suprasegmental characteristics. The merging and splitting of the move types to optimise intonation similarity is a possible way to improve move recognition (e.g. merging *align* and *check*). This would only be of use, however, if the resulting language models improve the word error rate.

8. ACKNOWLEDGEMENT

Helen Wright holds an EPSRC studentship (Award Ref. No. 96307159) and would like to acknowledge the assistance of Paul Taylor, Simon King and Stephen Isard.

9. REFERENCES

1. 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.
2. L. Breiman, J. Friedman, and R. Olshen. *Classification and Regression Trees*. Wadsworth and Brooks, Pacific Grove, CA, 1984.
3. Jean Carletta, A. Isard, S. Isard, J. Kowtko, A. A. Newlands, G. Doherty-Sneddon, and A. Anderson. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23:13–31, 1997.
4. S. King. *Using Information Above the Word Level for Automatic Speech Recognition*. PhD thesis, University of Edinburgh, 1998.
5. Jacqueline C. Kowtko. *The Function of Intonation in Task Oriented Dialogue*. PhD thesis, University of Edinburgh, 1996.
6. R. Power. The organization of purposeful dialogues. *Linguistics*, 17:107–152, 1979.
7. Elizabeth Shriberg, Paul Taylor, Rebecca Bates, Andreas Stolcke, Klaus Ries, Daniel Jurafsky, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech (in press)*, 1998.
8. Paul A. Taylor. Analysis and synthesis of intonation using the tilt model. *Proceedings of ICSLP88*, 1998.
9. Paul A. Taylor, S. King, S. D. Isard, and H. Wright. Intonation and dialogue context as constraints for speech recognition. *Language and Speech (in press)*, 1998.
10. University of Stuttgart. *Stuttgart Neural Network Simulator Manual*, 1996.
11. Steve Young, Joop Jansen, Julian Odell, Dave Ollason, and Phil Woodland. *HTK manual*. Entropic, 1996.