Research Proposal

1. Introduction

In this project we shall investigate variation in English intonation, insofar as it is represented in the IViE corpus. We shall employ techniques from experimental phonetics and corpus-based computational linguistics to characterise variation between dialects, speaking styles, men and women, and individual differences or 'habits'. In the first half of the project, we shall conduct statistical examinations of that variation. In the second half we shall construct a computational model of the mapping between intonation labels and the frequency of vocal cord vibration, f_0, and how that mapping differs according to dialects, style, gender, and individuals. We will also model the sequential structure of the intonation transcriptions and of the f_0 contours.

1.1. Prior work on the IViE corpus

The IViE corpus of intonational variation in the British Isles (Grabe *et al.* 2000, Grabe and Post 2002) was collected by Grabe, Nolan and collaborators on a previous ESRC project. It includes recordings of nine urban dialects: Caribbean English spoken in London, Punjabi English spoken in Bradford, and the English of Belfast, Cambridge, Dublin, Cardiff, Leeds, Liverpool, and Newcastle. 54 male and 54 female 16-year-olds were recorded in an identical set of tasks, eliciting five speaking styles: read sentences, a read passage, re-telling (i.e. free oral paraphrase) of the read passage, problem-solving conversation (a map task) and free conversation about a given subject. The completed corpus contains 36 hours of speech.

A one-hour balanced subset of the recordings of all five speech styles has been transcribed using the IViE transcription system, an extension and adaptation of the ToBI standard for intonation transcription, allowing for directly comparable analyses of several dialects of one language in one notation. This system offers additional levels of transcription, allowing for separate records of phonetic and phonological characteristics of intonation. For two of the styles, we have a further 2½ hours of transcribed data from seven dialects, spoken by 3 male and 3 female speakers per dialect.

A typical example of the data contained in the IViE corpus is shown in Figure 1. Nine kinds of data are shown, from top to bottom: 1) the acoustic waveform of the sentence "you remembered the lilies", read in isolation by a female speaker from Belfast; 2) *rms*, the short-time root mean square amplitude, which is related to the loudness of the speech; 3) *prob_voice*, an (accurate) estimate of the probability of voicing; 4) f_0, an estimate of the fundamental frequency of the voice; 5) a phonological transcription of tone targets, time-aligned to the acoustic data; 6) a systematic-phonetic transcription of the intonation, made according to the IViE transcription conventions; 7) a transcription showing the location of prominent syllables; 8) a word-by-word transcription in ordinary orthography, time-aligned to the end of each word in the audio recording. The ninth parameter is time, shown on the horizontal axis.

Most of the project was devoted to collection and transcription of recordings, but a number of analyses were conducted as the project progressed, leading to a series of insights on variation in English intonation in the British Isles (Grabe 2002). First, Grabe showed that dialect is a significant variable in intonation typology. A comparison of data from the IViE corpus with data from earlier publications shows that the intonation systems of dialects of one and the same language can be as different as the intonation systems of different languages. For instance, in read speech, the phonological structure of Cambridge English intonation is closer to that of dialects of German spoken in Northern Germany than to that of Belfast English. Similarly, the intonation of Leeds English exhibits a phonetic realisation effect that has also been observed in Northern Germany, but not in Cambridge English.

The second major result involves the extent of intonational variation in the British Isles. The data provide evidence not only of differences between dialects but also of considerable variation across speakers. In a particular dialect and in a particular context, there is not just one appropriate intonation contour. Speakers have a repertoire of appropriate contours, and some contours are used more frequently than others (Figure 2). This finding is new: previous descriptions of English intonation, especially since Kingdon (1958: 208–246), have implicitly or explicitly been based on rules of the type "for a given text, in context X (or with meaning X), use intonation contour Y". The IViE data show that a more adequate description of intonation needs to be based on rules such as:

For a given text, in dialect A in context B, use intonation contour C1 (with probability X) or
intonation contour C2 (with probability Y) or
intonation contour C3 (with probability Z)

This finding presents a serious challenge to intonation researchers. We need to consider how models of intonation can accommodate the variation that we have uncovered. Grabe's work also shows that intonation typology needs comparable data from several speakers of a dialect, several dialects of one language, and a number of languages.

This project will build on Grabe's (2002) findings. That study was limited to the effect of dialect in read sentences and to phonological transcriptions of intonation. We will extend this to take account of the effects of speaking style, gender and speaker habit. Finally, we will examine differences in the mapping between the phonological and the phonetic transcriptions in the corpus and the acoustic signal.

1.2. Prior work in speech technology

The present proposal builds on the IViE project, by exploiting techniques from speech technology research to develop a precise model of intonation variation. Computational modelling of the mapping between the discrete symbols of intonation transcriptions and continuous f_0 contours is a task that is closely related to the extensive body of earlier work on "prosodic synthesis-by-rule" by Coleman and collaborators (e.g. Coleman 1992, Dirksen and Coleman 1997). In the present project, we shall follow Pierrehumbert (1980) in employing finite state automata to encode the intonational grammars of specific dialects, as in Pierrehumbert's finite state automaton for American English intonation (Figure 3). Finite state techniques are now a commonplace method for computational modelling of tone and intonation. In recent years, however, probabilistic extensions of finite state automata have attracted interest. An influential paper by Coleman and Pierrehumbert (1997) provided evidence for the psychological reality of finite-state probabilistic grammars in phonology, a result that has been supported in psychology experiments by Frisch *et al*. (2000) and Bailey and Hahn (2001). The influence of that paper is also shown in Dainora (2001), a University of Chicago PhD thesis that presented a probabilistic extension of Pierrehumbert's model of American English intonation (Figure 4), trained on the intonation data in the Boston University Radio News Corpus. This work is an example of the type of work to be conducted in the modelling part of the present proposal.

In the 1970's and 1980's, synthesis-by-rule methods based on linguistic theories of prosody were popular (e.g. Allen *et al*. 1987, ch. 10, Pierrehumbert 1981). The principal strength of this approach is that the relationship between linguistic units and f_0 contours is expressed fairly transparently, in a way that can be understood by humans. However, this method suffers from a number of well-known weaknesses: i) rule sets are hard to debug, because interactions between rules are hard to foresee, and may produce unwanted, undesirable outcomes. ii) Rule systems are fragile: although they may be applied to novel texts, they contain unknown and unexpected gaps in their coverage, so that they frequently fail to work outside the research environment. iii) Effectively, only a single speaker and dialect are modelled by a given rule set. To model different voices and dialects, any or all of the rules might have to be altered by hand, which is almost as hard as starting from the beginning again with the construction of a completely new model.

Because of these difficulties, there was an explosion of interest in the 1990's in data-oriented modelling techniques, sometimes embodying machine learning or other "trainable" approaches (e.g. neural networks, Hidden Markov Models; see Taylor *et al.* 1998, Dusterhoff 2000). For developers of speech synthesis and recognition applications, methods such as these are very attractive, for several reasons: i) they apparently require little or no involvement of human experts; ii) they are mathematically well-understood techniques with known abilities to generalise over novel input and gaps in the training data; iii) they can be more easily extended to new speakers with different dialects or speaking styles. iv) the quality of synthetic speech generated by such techniques is often better than that of synthesis-by-rule.

Four such modelling methods have been employed: i) "neural networks", ii) Hidden Markov Models (HMM's), iii) regression models (e.g. CART, general linear models), and iv) curve-fitting models. Neural networks and HMM's suffer from the problem of being linguistically rather unrevealing. It is difficult to extract and interpret details of the mapping between intonation transcriptions to f_0 contours in a way that is meaningful to intonation researchers. Curve-fitting methods such as Fujisaki's model (e.g. Fujisaki 1997) focus on mathematically describing f_0 contours with as few parameters as possible. While this strategy is sensible in certain contexts with no firm theory of intonational phonology, the relationship between curve parameters and linguistic descriptions of intonation is opaque. In this project, though, we desire methods that are capable of relating two *given*, distinct types of representation: intonation transcriptions and acoustic parameters such as f_0. These methods will allow us to combine state-of-the-art analysis and modelling techniques with linguistic knowledge.

2. Detailed research questions

Like other acoustic parameters, f_0 contours exhibit variation due to many factors, including:

i) the linguistic structure of intonation (i.e. intonational phrasing) and sequences of tones (transcriptions);
ii) intrinsic f_0 variations due to the vowels and consonants making up the utterance;
iii) speaker-specific factors (e.g. speaker habits and individual voice characteristics);
iv) other pragmatic intentions (e.g. lowering or raising the loudness of the voice);
v) dialect;
vi) physical limitations on voice dynamics;
vii) speaking style;
viii) context of use;
ix) the speaker's sex.

Consequently, each observed f_0 value must be related to a variety of factors. We shall address the following questions:

1) How do these factors, especially i), the linguistic structure of intonation, v), dialect, vii), speaking style and viii), context of use, and their interactions, relate to f_0?
2) To what extent, and in what ways, do iii), speaker-specific factors, and ix), gender differences, play a role in English intonation?
3) How can intonation variation be properly represented and modelled, at the linguistic level (e.g. variation in intonation transcriptions), at the acoustic level (variation in f_0 contours) and variation in the mapping between these two levels.

Additional, more particular questions are given in section 4.1 below.

3. Data

As stated above, no new data will be collected in this project. Instead, we shall analyze the data already collected in the IViE project in much greater depth than in the pilot analyses of earlier work. We shall employ two balanced subsets of the IViE corpus covering each dialect and style and both genders. One subset, of one hour's duration, covers all five styles and seven dialects, spoken by one male and one female speaker of each dialect. The second subset, of 2½ hours' duration, covers three male and three female speakers of seven dialects, in only two styles. 90% of each of these subsets will be used for subsequent analysis and modelling: the remaining 10% will be kept aside for use in later evaluations of the model.

4. Framework and methods for analysis

Our model will have the following design characteristics:

1) The model shall account for variation due to dialect, speaking style, sex and individual speaker habits. Other models focus on a single dialect (usually RP or "General American"), a single speaking style (careful, read speech), and they usually do not take individual differences into account.

2) The model should reveal the nature of the mapping between f_0 contours and intonation categories in a way that is meaningful to researchers. In contrast, many other models of intonation are developed with the primary goal of generating f_0 contours accurately, by whatever means, without revealing to human researchers the nature of the mapping.

3) The model shall preferably have a relational, bi-directional character, rather than being oriented to generation or analysis in particular.

4) Rather than seizing on a particular, favoured modelling technique, in the first year we shall employ general statistical techniques to determine empirically the structure of the data. Then, to model the sequential structure of the data, we shall infer probabilistic finite state automata from the IViE corpus. To characterise the mapping between intonation transcriptions and f_0 contours, we shall employ probabilistic finite-state transducers, to model observed intonational melodies and the mapping between intonation transcriptions and f_0 values. We shall also examine the use of sums-of-products models (van Santen 1993) to model unseen events. This part of our model will be similar to the model of segmental duration proposed by van Santen and Olive (1990).

5) We shall not attempt to model the *assignment* of intonation symbols to a text: that is a very difficult problem, involving, *inter alia*, computation of discourse structure and some aspects of meaning.

4.1. Workplan

Phase 1. Statistical exploration of the data. Before trying to model the corpus sample, we shall obtain informative statistics concerning the variation in f_0 contours between speakers, dialects and speaking styles, as exhibited in the entire corpus. We begin by analyzing variation in the intonation transcriptions. Then we shall move on to an analysis of acoustic variation.

a) Statistical analysis of the intonation transcriptions. To determine what patterns occur with what frequency for each dialect, style, gender and speaker, we shall examine the variation in the initial, medial and final (nuclear) positions of intonation melodies, and in read passage data, the effect of position in the narrative (whether paragraph initial, medial or final). We shall determine whether individual speakers have frequent patterns that cut across all styles, to assess speaker habits, and we shall establish the most frequent accent patterns for each speaker, gender, style and dialect.

For the read passages, we shall then examine the correlation between tunes and their context of use (i.e. position in the passage) across dialects. The purpose of this task is to commence the investigation of semantic differences in intonation across dialects and sex. For example, do instances of the same pattern occur in different dialects and both sexes? If so, are they used for different purposes? This will be assessed by an inspection of the contexts in which those patterns are used, since we cannot access speaker's purposes directly. Therefore, we shall take "context of use" to indicate differences in the meaning or function of different intonation contours. We shall ask whether similar patterns are used for different purposes in the map task data, or in different contexts in the read passages.

b) Statistical analysis of f_0 *variation.* If the f_0 variation is not normally distributed (from prior work we expect that it will not be) we shall assess various standard transformations to see how the data may best be normalised. By normalizing f_0 across subjects, we will be able to combine data from several speakers within a given dialect or speaking style. Outliers of the f_0 distributions are often due to limitations of the f_0 tracking software: identification of outliers will also enable us to disregard those data points when constructing the model in phase 2. As a result of this task, we shall be able to describe f_0 variation due to dialect, style, gender, intonation accent type, and position in the intonation phrase.

We will present the results of this phase in two papers: one on variation in the intonation melodies, the other on acoustic variation in f_0 and (possibly) RMS amplitude.

Phase 2. Probabilistic finite-state modelling. In this phase, we examine the way in which intonation variation varies with the sequential structure of intonation transcriptions (and in the acoustics, with time). There are three main tasks. First, we shall construct probabilistic finite-state models of the intonation patterns alone. Second, we shall time-align similar f_0 contours, using dynamic time warping. Third, we shall construct probabilistic finite-state transducers, to model the mapping between intonation transcriptions and f_0 contours.

For the first task, we shall distil lists of intonation patterns observed in the database to construct finite state automata. We shall examine whether the differences between dialects and styles warrant different grammars (i.e., whether they are systemic differences), or whether the differences are mainly statistical. We shall examine the correlation between different phonotactic patterns (i.e. intonation morphemes) and the different dialects and styles.

We shall use dynamic time warping (DTW) to calculate 'prototypical' f_0 contours for each dialect and style, together with measures of the variability at different points in the contour. This will be of particular value to work that requires 'prototypes', such as e.g. language teaching, as well as in general presentations of our results. Furthermore, by warping tokens of one dialect or style onto another, we shall be able to determine where the dialects or styles are most different, by analyzing differences in the warp functions. It will also be necessary to synchronise the utterances in time in order to combine the state sequences of multiple tokens of utterances spoken by different speakers on different occasions (most likely with different speaking rates). We shall employ DTW for this too. Without DTW, one cannot properly estimate the variance and means of the accent realisations.

We shall then construct probabilistic finite-state models of the accent–f_0 mappings by clustering the data in the corpus. To illustrate this strategy, consider the sentence in Figure 1. By tabulating the (time-warped) f_0, *prob_voice*, and tone target tiers, this utterance may be segmented into twelve intervals and two endpoints. From these, a small finite state transducer can be straightforwardly constructed (Figure 5). Finite state transducers for individual utterances will be combined with those of other speakers within each dialect and style.

At that point, a tally can be made of the frequency of each transition, from which transition probabilities can subsequently be estimated. The f_0 values associated with each transition of the FST amount to a tabulation of the database. Equivalent transitions (that is, all the transitions from one state to the same next state) shall be replaced by a single transition associated with an f_0 probability distribution, covering the several observed instances.

Since our goal is to understand the data in the corpus, rather than to build an intonation synthesiser, we shall initially only include states and transitions exhibited in the data in the finite-state transducer. As we construct FSTs for each dialect, style and gender, we shall compare them to determine where, how, and to what extent they are different. Finally, we shall measure the goodness-of-fit of models for each dialect and style on held-out data (10% of the balanced database) and determine the extent of the problem of dealing with unseen data.

The anticipated timetable of research is summarised as follows:

Phase	Months from start	RA 1 (Grabe)	RA 2
1	3	Statistical analysis of intonation transcriptions in the corpus	
	6	Study phonological variation in transcriptions	Study and write paper on acoustic variation
	8	Write paper on phonological variation in transcriptions	
	9		Definition of probabilistic grammars
2	11	Definition of probabilistic grammars	
	13	Documentation of probabilistic grammars	
	15	Time-warping of f_0 contours of selected sample	
	17	Combine finite state automata with acoustic models to make finite state transducers	
	19	Compare FST's for different dialects, styles, gender	FST construction and evaluation
	21	Write final paper, final report, public leaflet	
3	24		Mount results on IViE web site

5. Conclusion of project

The final written paper of the project will be a description of intonational variation in several English dialects, with reference to style and gender variation. It will include simple finite-state descriptions of the main patterns, and prototypical f_0 contours. As well as this paper and the final report for ESRC, we propose to update the IViE website with the addition of outputs of this project. Also, we shall publish a pamphlet on our results for dissemination to colleagues in the speech therapy community, who have expressed their eagerness for this deliverable. The content of this pamphlet will also be mounted on the project website so that it can be accessed by anyone else who may be interested, e.g. English language teachers.

Annex

References

Allen, J., M. S. Hunnicutt and D. Klatt (1987) *From text to speech: The MITalk system.* Cambridge University Press.

Bailey, T. M. and U. Hahn (2001) Determinants of wordlikeness: phonotactics or lexical neighbourhoods? *Journal of Memory and Language* **44**, 568–591.

Coleman, J. S. 1992. "Synthesis-by-rule" without segments or rewrite rules. In G. Bailly, C. Benoit and T. R. Sawallis, eds. *Talking Machines: Theories, Models, and Designs*. Amsterdam: Elsevier. 43–60.

Coleman, J. S. and J. B. Pierrehumbert (1997) Stochastic Phonological Grammars and Acceptability. In *Computational Phonology. Third Meeting of the ACL Special Interest Group in Computational Phonology*. Somerset, NJ: Association for Computational Linguistics. 49–56.

Dainora, A. (2001) *An empirically based probabilistic model of intonation in American English*. PhD dissertation, University of Chicago.

Dirksen, A. and J. S. Coleman (1997) All-Prosodic Synthesis Architecture. In J. P. H. van Santen, R. W. Sproat, J. P. Olive and J. Hirschberg, eds. *Progress in Speech Synthesis*. New York: Springer-Verlag. 91–108.

Dusterhoff, K. E. (2000) *Synthesizing Fundamental Frequency Using Models Automatically Trained from Data*. PhD thesis, University of Edinburgh.

Frisch, S. A., N. R. Large and D. B. Pisoni (2000) Perception of wordlikeness: effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* **42**, 481–496.

Fujisaki, H. (1997) Prosody, models, and spontaneous speech. In Y. Sagisaka, N. Campbell and N. Higuchi, eds. *Computing Prosody: Computational models for processing spontaneous speech.* New York: Springer-Verlag. 27–42.

Grabe, E. (2002) Variation adds to prosodic typology. In B. Bel and I. Marlin (eds), *Proceedings of the Speech Prosody 2002 Conference*, 11–13 April 2002, Aix-en-Provence: Laboratoire Parole et Langage. 127–132.

Grabe, E. and Post, B. (2002). Intonational Variation in English. In B. Bel and I. Marlin (eds), *Proceedings of the Speech Prosody 2002 Conference*. 11–13 April 2002, Aix-en-Provence: Laboratoire Parole et Langage. 343–346.

Grabe, E., Post, B., Nolan, F., and Farrar, K. (2000). Pitch accent realisation in four varieties of British English. *Journal of Phonetics* **28**, 161–186.

Kingdon, R. (1958) *The groundwork of English intonation.* London: Longman.

Pierrehumbert, J. B. (1980) *The phonology and phonetics of English intonation*. PhD dissertation, MIT.

Pierrehumbert, J. B. (1981) Synthesizing intonation. *Journal of the Acoustical Society of America* **70**, 985–995.

Taylor, P., S. King, S. Isard and H. Wright (1998) Intonation and Dialog Context as Constraints for Speech Recognition. *Language and Speech* **41**, 493–512.

van Santen, J. P. H. (1993) Exploring *N*-way tables with sums-of-products models. *Journal of Mathematical Psychology* **37**, 327–371.

van Santen, J. P. H. and J. P. Olive (1990) The analysis of contextual effects on segmental duration. *Computer Speech and Language* **4**, 359–390.

Figures

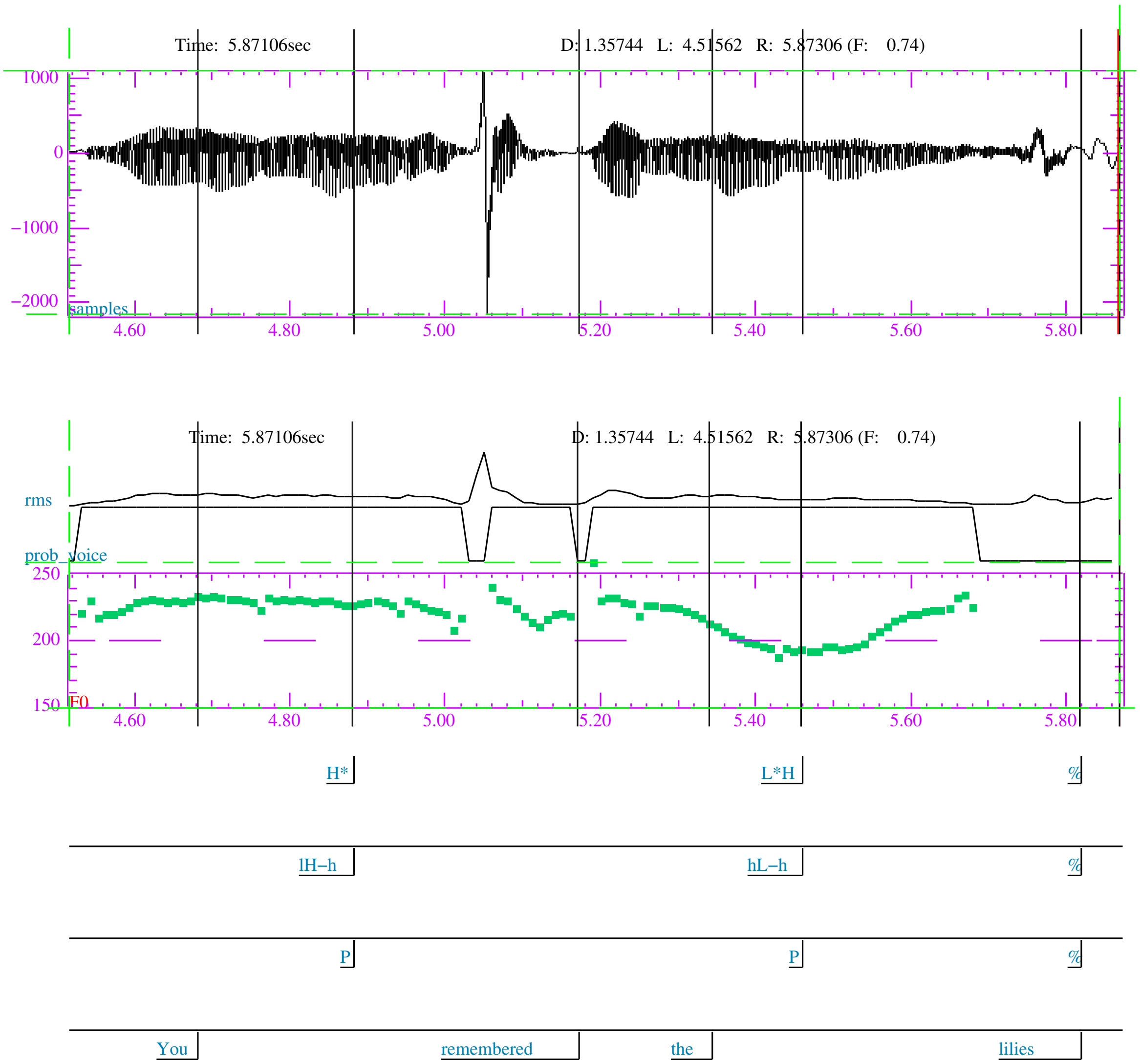

Figure 1. A typical example of the data contained in the IViE corpus. From top to bottom: 1) acoustic waveform of the sentence "you remembered the lilies", read in isolation by a female speaker from Belfast; 2) *rms*, the short-time root mean square amplitude; 3) *prob_voice*, an (accurate) estimate of the probability of voicing; 4) f_0, an estimate of the fundamental frequency of the voice; 5) a phonological transcription of tone targets, time-aligned to the acoustic data; 6) a systematic-phonetic transcription of the intonation, made according to the IViE transcription conventions; 7) a transcription showing the location of prominent syllables; 8) a word-by-word transcription in ordinary orthography, time-aligned to the end of each word in the audio recording. Time is shown on the horizontal axis.

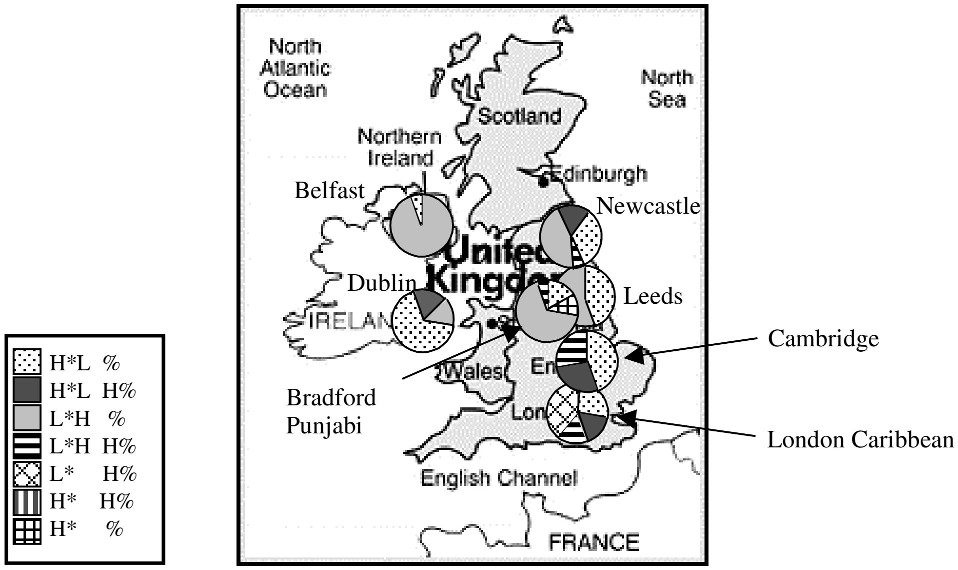

Figure 2*:* Intonation patterns in yes/no questions in seven dialects of English (e.g. *May I lean on the railings?*), from Grabe and Post (2002). This figure is based on the read sentence data. It illustrates intonational variation in the British Isles. Firstly, the figure shows broad geographical differences in the distribution of nuclear accents. Rise-plateau patterns (L*H %), for instance, are found in northern cities, but not in London or Cambridge. Secondly, it shows that dialects cannot be distinguished from each other on the basis of a single, dialect-specific intonation contour. In each dialect, more than one contour is appropriate, and dialects differ because speakers have different dialect-specific choices.

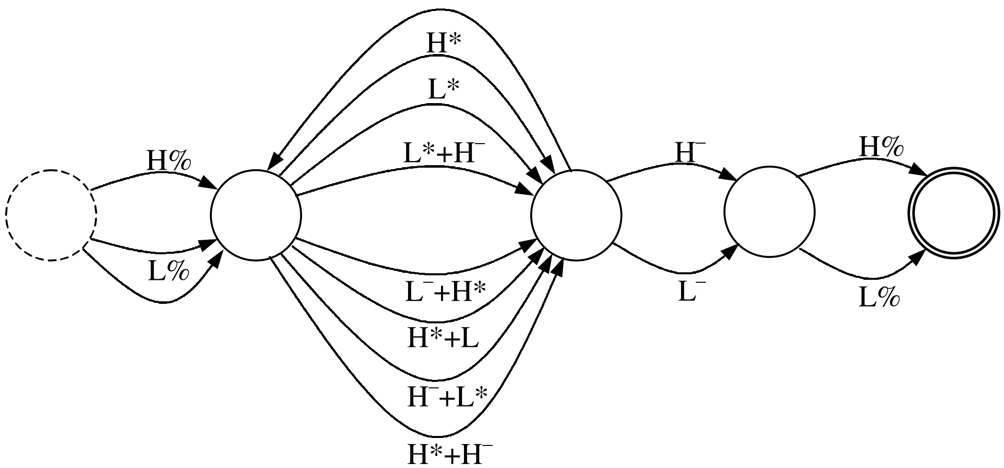

Figure 3. Finite state automaton defining the possible intonational transcriptions of American English, proposed by Pierrehumbert (1980).

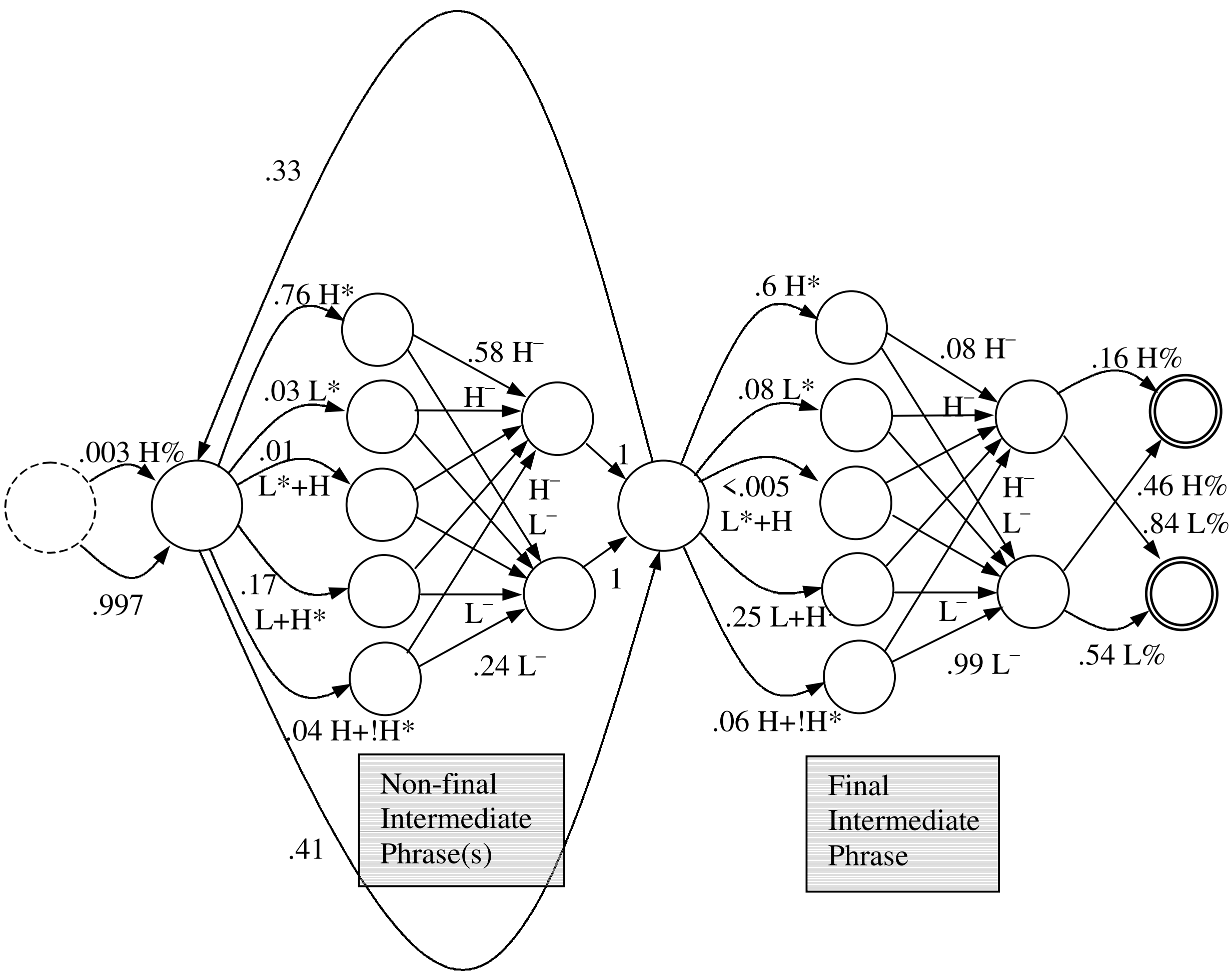

Figure 4. Probabilistic finite state automaton for American English intonation, adapted from Dainora (2001).

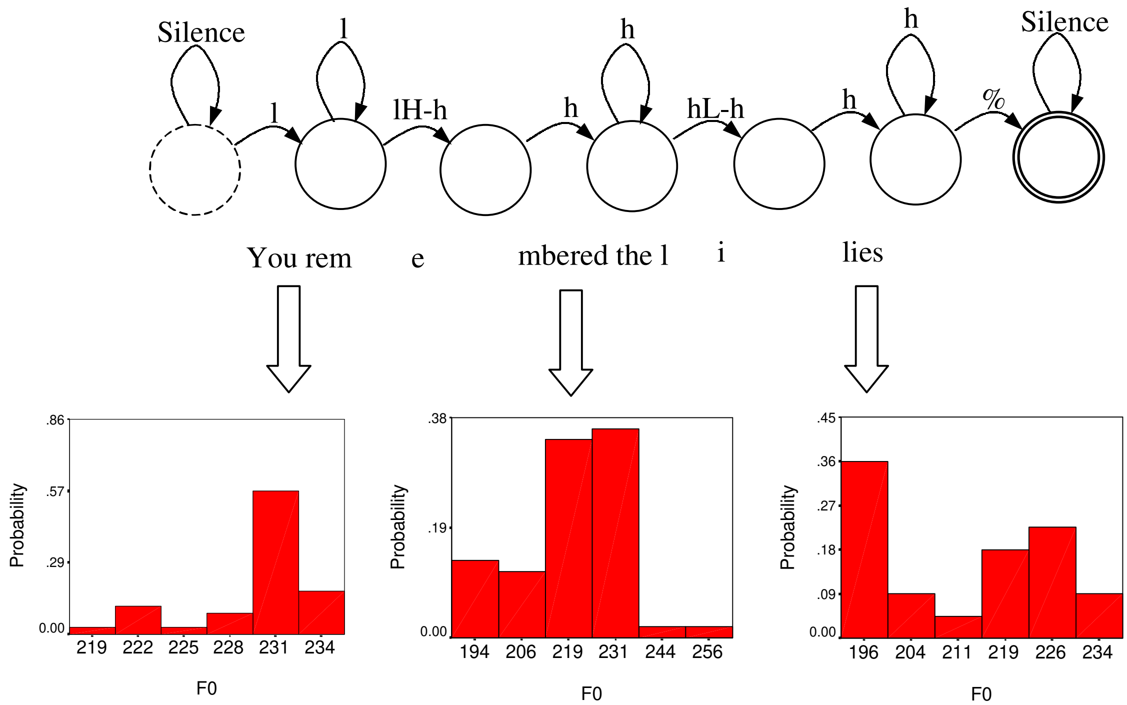

Figure 5. Probabilistic finite state transducer constructed from Figure1, encoding the mapping between the phonetic transcription of intonation and observed f_0 values. In this example, the f_0 probability distributions shown in the histograms are based on data from a single utterance. In the project, distributions will be calculated across the corpus. With this type of model, it will be possible to determine, for a given transcription, dialect, style, gender and speaker the most likely f_0 contour, as well as possible but less common realisations.