Automatic Intonation Analysis Using Acoustic Data

Kurt Dusterhoff

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

ABSTRACT

In a research world where many human-hours are spent labelling, segmenting, checking, and rechecking various levels of linguistic information, it is obvious that automatic analysis can lower the costs (in time as well as funding) of linguistic annotation. More importantly, automatic speech analysis coupled with automatic speech generation allows human-computer interaction to advance towards spoken dialogue. Automatic intonation analysis can aid this advance in both the speaker and hearer roles of computational dialogue. Real-time intonation analysis can enable the use of intonational cues in speech recognition and understanding tasks. Auto-analysis of developmental speech databases allows researchers to easily expand the range of data which they model for intonation generation.

This paper presents a series of experiments which test the use of acoustic data in the automatic detection of Tilt intonation events. A set of speaker-dependent HMMs is used to detect accents, boundaries, connections and silences. A base result is obtained, following Taylor [8], by training the models using fundamental frequency and RMS energy. These base figures are then compared to a number of experiments which augment the F0 and energy data with cepstral coefficient data. In all cases, both the first and second derivative of each feature are included. The best results show a relative error reduction of 12% over the baseline.

1. INTRODUCTION

The body of research into manual and automatic intonation analysis systems and techniques has been growing rapidly in the last few years. It is notable that only two moderately successful automatic intonation analysis systems relate to the current trends in intonation description. Ostendorf and Ross [5] use a system which works with syllables to determine pitch accent location and type. Taylor [8] takes a waveform and determines pitch accent location from acoustic information derived from the waveform. Both of these systems are fully functional, but neither is as successful as one might like for use in speech recognition/understanding systems or as a database labelling tool.

The goal of the intonation analysis research detailed in this paper is to create a system which can automatically label speech with intonation information. Spoken language understanding systems can benefit from the structural and pragmatic information which intonation often conveys. Current trends in speech processing have increased the need for large corpora. Stochastic speech synthesis methods, including those used for intonation modelling, require a great deal of data to be effective. While word recognition is seen to have reached a level suitable for application, automatic intonation analysis is in its infancy. Manually labelling speech databases for intonation is recognized as difficult and time consuming. Automatic labelling can decrease both time and funds spent on building the databases from which theoretical models and viable applications can be built.

Automatic intonation analysis methods which require other types of prior analysis have achieved some success in the past, provided that the prior steps are highly successful. However, error introduced by initial analyses of syllable, segment, and prosodic phrase boundaries can render useless a model which requires them. In addition to the potential problems from prior analyses, such systems are less likely to assist in further speech recognition tasks, such as text disambiguation, as the output of such systems is required prior to intonation analysis. Therefore, a model which does not rely on any possibly inaccurate prior linguistic interpretation of the speech signal should provide an improvement in applicability and quality over other types of models.

This paper discusses an extension of the idea behind Taylor's intonation analysis method. This extension is to expand the acoustic data used for intonation analysis to include information about cepstral coefficient data. To place this research into a wider context, basic intonation analysis problems are presented.

2. INTONATION ANALYSIS

Intonation analysis generally involves three basic tasks: event detection, event identification, and event-syllable association. Detection of intonation events involves determining where, in the speech signal, accent and boundary events are located. Identification of intonation events consists of giving names to each event. In the Tilt model, for example, identification involves determining whether an event is an accent, a boundary, or perhaps a combination of both. Using the ToBI model, the process involves not only determining whether the event is an accent or boundary, but what the tones are that make up the event. The third task, association, is the act of linking an event with a portion of linguistic text (e.g. syllable nucleus, demi-syllable, syllable, word, or phrase). This paper is concerned with event detection. However, the model-building process involves first building models of individual event types, and then using all of the smaller models to detect events in novel speech. Therefore, the detection process utilizes models of specific event types, but the detection evaluation counts two different event types as being equivalent. Details of the use of this evaluation technique are discussed below.

3. EVALUATION

The output of the various experiments is evaluated in terms of three basic measures: percent of detected labels which are correct, accuracy (correct minus percent of detected labels which are incorrect), and error (100% minus accuracy). While seemingly simple, this evaluation scheme requires a definition of correctness. With intonation, correctness is, to some extent, in the ear of the listener. For the purposes of this paper, a detected label is deemed correct when it overlaps an original event by at least 50%. This loose definition allows for the equivalent of two human labellers disagreeing on the exact location of an accent within a word, while agreeing that the word is accented.

As mentioned previously, the task being carried out in this study is primarily one of event detection. However, there is a degree of event identification involved as well. Each event type has a Markov model built for it. Events are detected on the basis of fitting any one of the event models. Therefore, during evaluation, an accent in the original label file and a detected falling boundary, if fulfilling the timing requirement for correctness, result in a correct event detection.

The principle reason that this loose definition of correct matching is acceptable is that, in the Tilt intonation model, events of all types are described using the same parameter set. Therefore, event types are really a convenience for the human interpreter, and are not necessarily important for computing applications. Additionally, studies have shown that humans will agree to a greater extent on the location of an intonation event than on its type [3], [6].

4. DATA

The research is primarily based on 45 minutes of radio news broadcast from the Boston University Radio Corpus [4], speaker F2B (over 5000 intonation events). Other corpora examined are three databases spoken by the author (male American English speaker). Of these three, one is a TIMIT-style database (KDT - 2000+ events), one is a series of weather-related sentences (KDW - 2400+ events), and the third is a museum guide (KDS - 3200+ events). Each corpus has been hand-labelled with Tilt intonation labels. The intonation event inventory for this study is accents, rising boundaries, falling boundaries, and concatenated accents and rise/fall boundaries (this represents an extended inventory of the Tilt model).

The acoustic information was extracted using the following methods. In each case, the fundamental frequency was derived using Taylor's Intonation Contour Detection Algorithm [9] which provides a smoothed, interpolated F0 trace. The smoothing algorithm uses windows of 105ms (first pass) and 35ms (second pass) to remove outlying points, but to leave behind as much contiguous data as possible (thereby providing as much micro-intonation information as possible while removing isolated outlying F0 points). The Mel Frequency Cepstral Coefficients were calculated using the HCopy function of the Entropic HTK package. The energy information was extracted from Entropic's get_f0 output. The F0 and energy values were normalized on a scale of -1 to 1 for each database individually, based on the mean and standard deviation of the respective values.

5. METHODOLOGY

The Hidden Markov Models used in these experiments are created using Entropic's Hidden Markov Model Toolkit [10]. In each case, unless otherwise noted, five-state, left-to-right HMMs are used. The states roughly represent the beginning, rise, peak, fall, and end of a pitch event. Transitions exist from state to state serially, as well as from beginning to peak and peak to end. By allowing the skipping of states, the models match a conceptual model where a pitch event is rise-fall, rise, or fall (e.g. the Tilt intonation model). One experiment with a four-state model (conceptually leaving out the peak state) was undertaken. No noticeable difference was found between the four- and five-state models, with the relative error rates separated by less than one hundredth of a percentage point.

The models were trained on 70% of the speech data, and tested on 30%, except in the case of the F2B database,

where the test set contains 20% of the data and 10% was held out for blind testing at the end of all experiments.

All of the tests were constrained by a bigram/unigram grammar which was built from the corpus being tested. Models were trained using odd-numbers of Gaussian components from 1 to 29. Scores were obtained for each set of models. Only the best results of each database are reported here. For each database, initial evaluation of a grammar scaling factor was undertaken to determine the general range of productive grammar weighting values. The weights tested ranged from 3 to 20 (where 0 is no reference to the grammar at all).

A similar set of tests examined the use of an external transition weighting (to weight from the command line the transition probabilities). A negative value lowers the transition probability (which reduces insertions), while a positive value raises the transition probability (which increases insertions). Values were tested from -60 to 30 at five-point intervals.

Most of the scores which are reported in this paper were achieved with constraints optimized on the test data, for speed and efficiency. However, the HMMs and optimized constraints which received the best scores were also used to automatically label the blind (held-out) set once all other experiments were complete. This score is comparable to the score received for the test set, as is discussed below.

6. EXPERIMENTS

A portion of Taylor's study examines event detection of the F2B data, and is the basis of the baseline experiment. Taylor built models of intonation event types using F0 and RMS energy in various forms. The portion of his research which relates to this study used normalized F0 and RMS energy, together with the first and second derivatives of each feature. The results of the experiments which are relevant to this chapter are 79% of detected events correct, and 59% accurate (error of 41%). Taylor's use of normalized values stems from his desire to create a speaker-independent analyzer. Both normalized and non-normalized values were used in the research discussed here. First, non-normalized F0 and RMS energy were modelled, with results (Base 1) in Table 1 of 78% correct and 61% accuracy (error of 39%).

	Correct	Accuracy	Error
Taylor	79%	59%	41%
Base 1	78%	61%	39%
Base 2	78%	59%	41%

Table 1: Comparison of baseline results

As these results were reasonably close to Taylor's, normalized F0 and RMS energy were modelled in order to provide a direct comparison to [8]. The results of this experiment (Base 2) were 78% correct and 59% accuracy (error of 41%). The close similarity of these results allows for a reasonable comparison between any results in this chapter and [8].

The use of cepstral coefficients reflects some of the experimental findings in the literature. Spectral tilt and general formant information are represented in cepstra. Campbell and Beckman [1], among others (e.g. [7] [2]) , have provided support for links between spectral tilt and the existence of pitch events. A variety of formant information can provide useful information about the type of segments associated with a given pitch event. Such information should be useful in lowering the number of pitch movements which are incorrectly analysed as intonation events.

Table 2 shows results from experiments with non-normalized F0 and all thirteen MFCC (all data for all experiments includes first and second derivatives).

Accuracy	Relative Error
67.5%	-15%

Table 2: Accuracy and Relative Error Compared to Baseline of experiments using Mel Frequency Cepstral Coefficients and F0

Instead of simple F0 information for the next series of experiments, normalized F0 was used, in order to allow direct comparison between this work and previous work [8]. Table 3 shows the results for the two best weightings for the normalized F0 and MFCC experiments. The smaller weighting produces similar, but slightly better error reduction.

Weight	Weighted with F0	Relative Error to Baseline
0.8	63%	-10%
0.6	64%	-12%

Table 3: Error of experiments using Mel Frequency Cepstral Coefficients to augment Normalized F0 and energy, with relative error

The relative error reduction of the MFCC experiments is encouraging, but it could also be misleading. The purpose of this research is partly to remove insertion errors from automatic detection. The manner in which error is calculated allows for an error reduction without a decrease in insertions (by improving correct detection). While an increase in correct analyses is beneficial, it is partially a by-product

of the drive for lower insertions. Therefore, an investigation of all three evaluation metrics is useful to determine whether using MFCCs to reduce insertions has been a success.

	Correct	Accuracy	Error
Base 1	78%	61%	39%
Non-normalized MFCC	84%	67%	33%
Taylor	79%	59%	41%
Base 2	78%	59%	41%
Normalized MFCC	80%	64%	36%

Table 4: Comparison of results to baselines and Taylor 1998

Table 4 shows a comparison of the MFCC experiments with the respective baselines and [8]. As accuracy is correct minus the percentage of detections which are insertions (incorrect), it is important not only that the correct score rises, but also that the gap between correct and accuracy shrinks. The non-normalized experiment shows a rise in both correct and accuracy scores, resulting in a reduction of error. However, one may note that the relative percentage of insertions has remained the same (17 points). This means that the error reduction, while welcome, is not the result of reduced insertions. The results of the normalized data, in contrast, show both an improvement in correct identification and a reduction of insertions (from 19 points 16 points). While the normalized data does not show as large an improvement over Base 2 as the non-normalized data shows against Base 1, the improvement appears to be on a wider scale. With the experiments on F2B finished, the HMM set and constraints which produce the improvement over Base 2 (the normalized data) were used to automatically label the previously unseen 10% of the database. The resulting scores of 85% correct and 66% accuracy (34% error) show that the methodology was not overly biased towards the data used in optimizing the constraints. Table 5 shows how the manual and automatically recognized labels compare. These labels were taken from the unseen dataset.

7. DISCUSSION

The level of improvement which is achieved by adding cepstral information to the intonation analysis process indicates that acoustic data which reflects the type of segmental text associated with an intonation contour is useful for intonation analysis. In order to press this claim, three databases were tested in addition to F2B. As discussed above, each database is substantially smaller than F2B. Therefore, no blind set was held out for further use, primarily because it would consist of no more than a paragraph or two. Instead, the tests rely on the assumption gained from F2B that, given a reasonable sized database, the blind set will score similarly to the general test set.

Manual	Labels	Recognized	Labels
End Time	Label	End Time	Label
0.027	sil	0.040	sil
0.150	c		
0.358	a	0.510	a
0.618	c	0.590	c
0.863	a	0.850	a
1.020	c	0.980	c
1.270	a	1.240	afb
1.479	c	1.310	c

Table 5: Example manual and automatic label comparison

	Correct	Accuracy
Normalized Data		
F0 + Energy	71.08	56.21
F0 + MFCC (weight 0.8)	77.82	60.28
F0 + MFCC (weight 0.6)	75.96	59.93
Non-Norm		
F0 + Energy	71.31	56.44
F0 + MFCC (weight 0.8)	73.98	59.12
F0 + MFCC (weight 0.6)	74.1	59.7

Table 6: Analysis Results for Database KDS

Table 6 shows how KDS, the largest of these databases scored. The most notable aspect of these scores is that on all counts, they are considerably lower than those for F2B. The list of possible reasons for this difference is extensive. The most likely reason is that the database is 60% the size of F2B. This is born out by KDW giving no cohesive results. For this very small database, the process failed to result in HMMs capable of producing sensible label files at all.

This paper has shown that it is possible to improve upon previous methods of automatic intonation analysis without relying on interpretations of acoustic data (e.g. phone, syllable, word annotation). The methods described in this paper work exclusively with acoustic information which is readily available from the speech signal. The most important difference between this research and previous research which uses only acoustic data is this work presents a way of approaching the interaction between the supraglottal vocal tract and intonation. This paper began expressing a desire to increase the speed with which such data can become available. To this end, a method which provides better automatic intonation annotation than other comparable techniques was introduced. The limited success this method

achieved on a 45 minute speech database can be useful in developing the type of bootstrapped database growth that other areas of speech recognition encountered fifteen years ago.

REFERENCES

[1] N. Campbell and M. Beckman. Stress, prominence, and spectral tilt. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Intonation: Theory, models and applications*, pages 67–70. Athens, September 1997.

[2] J. Hirschberg and G. Ward. The influence of pitch range, duration, amplitude, and spectral features on the interpretation of the rise-fall-rise intonation contour in English. *Journal of Phonetics*, 20(2):241–251, 1992.

[3] C. Mayo, M. Aylett, and D.R. Ladd. Prosodic transcription of Glasgow English: an evaluation study of GlaToBI. In *Proceedings of ESCA Workshop on Intonation*, pages 231–234, Athens, Greece, 1997.

[4] M. Ostendorf, P. Price, and S. Shattuck-Huffnagel. The Boston University Radio News Corpus. Technical Report ECS-95-001, Electrical, Computer and Systems Engineering Department, Boston University, 1995.

[5] M. Ostendorf and K. Ross. A multi-level model for recognition of intonation labels. In Y. Sagisaka, N. Campbell, and N. Higuchi, editors, *Computing Prosody*, pages 291–308. Springer, 1997.

[6] 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 *Proc. ICSLP*, pages 867–870, 1992.

[7] A.M.C. Sluijter, V.J. van Heuven, and J.J.A. Pacilly. Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, 101(1):503–513, January 1997.

[8] P. Taylor. Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America*, forthcoming.

[9] P. Taylor, R. Caley, and A.W. Black. *The Edinburgh Speech Tools Library*. The Centre for Speech Technology Research, University of Edinburgh, 1.0.1 edition, 1998. http://www.cstr.ed.ac.uk/projects/speechtools.html.

[10] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *HTK manual*. Entropic, 1996.