An investigation into the effectiveness of sub-syllable acoustics in automatic intonation analysis.

Kurt Dusterhoff
kurt@cstr.ed.ac.uk

Abstract

This paper presents a series of experiments which test the use of sub-syllable acoustic data in the automatic detection of Tilt [Tayng] intonation events. A set of speaker-dependent HMMs is used to detect accents, boundaries, connections and silences. A base result is obtained, following Taylor, by training the models using fundamental frequency and RMS energy. A second baseline is obtained using normalized F0 and energy. These base figures are then compared to a number of experiments which augment the F0 and energy data with auto-correlation peak, zero-crossing, or cepstral coefficients. In all cases, both the first and second derivative of each feature are included. The baseline results of the normalized data are within one percentage point of those in Taylor on the same speaker, which supports the comparison of this study with Taylor's. The best results at present show a relative error reduction of 12% over the baseline.

1 Introduction

This paper presents current research into the effectiveness of low-level acoustic data in automatic intonation analysis. Current trends in speech processing have increased the need for large corpora from which to build models. Stochastic modelling of speech can be useful for adapting language models to new environments, dialects, or languages. However, a model which accounts for the wide variability in intonation structure and use requires a great deal of data for each speech situation to be modelled. Ideally, an automatic intonation analysis system increases both the amount of data which may be modelled and the speed with which the models are built.

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 creating linguistically annotated speech databases. However, as Mertens [Mer89] found, automatic intonation analysis is not simple. He attempted a construction of intonation annotation through analysis of prosodic, syllabic, and intonational structure. The complex interactions of the various levels seems to have made reporting his results impossible. Dusterhoff [Dus96] attempted to use an algorithm which finds points at which F0 changes direction as a means of determining pitch accent location. While the overall success of this method was better than anticipated, it suffered from a large number of spurious accent insertions. Similarly, Taylor [Tayng], while acheiving notable success in intonation event detection, finds that insertions pose a problem.

The current study, recognizing the need for accurate analysis, is centered on limiting insertion errors which mar otherwise successful results. Additionally, noting the difficulties that Mertens [Mer89] encountered in relying on the labelling of multiple layers of data before classification even begins, this study aims to use no more data than can be (relatively) easily extracted from the speech waveform.

2 Intonation Analysis

Intonation analysis generally involves three basic tasks: detection, identification, and placement. 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 the 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, placement, is the act of linking an event with a portion of linguistic text (e.g. syllable nucleus, demi-syllable, syllable, word, phrase). The task being undertaken is a combination of detection and identification.

The study is primarily one of event detection, in that all intonation event types are treated simply as events in the detection evaluation. 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 events, and therefore equivalent. Details of the use of this evaluation technique are discussed in Section 5.

3 Baseline Experiment

In order to determine the effectiveness of sub-syllable acoustic data on analysis, it was necessary to have a point for comparison. Taylor [Tayng] used Hidden Markov Models to model F0 and RMS energy for his intonation event detector. He also worked within the framework of his Tilt intonation model, which is the model planned for this study. Therefore, a portion of his study has been replicated to act as a baseline.

3.1 Data

The data for this study is 45 minutes of radio news broadcast from the Boston University Radio Corpus [OPSH95], speaker F2b. This 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 (these are the “major” events of the Tilt model).

3.2 Taylor 1998

The basis of comparison for this study is Taylor [Tayng]. A portion of Taylor’s study examines event detection of the F2b data. However, prior to the outset of this study, the hand-labelled events which he used were corrected as a requirement for use in unrelated experiments. Therefore, it is expected that, while very similar, the results of the replication experiments will differ somewhat from Taylor's results. In addition, it is unlikely that the training and testing data for this research is divided in the same manner as in Taylor's work.

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 relevent to this paper are 79% of detected events correct, and 59% accurate (error of 41%).

3.3 Replication of Taylor 1998

Two forms of Taylor's study were replicated in the process of creating baseline results. 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 [Tayng]. 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 paper and [Tayng].

4 Experimental Methodology

The Hidden Markov Models used in these experiments were created using Entropic's Hidden Markov Model Toolkit [YJO$^+$96]. The models were trained on 70% of the speech data, and tested on 30%. Initial tests were undertaken on models of F0 and RMS energy and one additional acoustic feature. Successful features were then used in a subsequent tests (again, only one feature was added to F0 and energy) where the models were trained with the added feature receiving various weights in relation to F0 and energy. Finally, tests were run on the most successful feature using normalized F0 and energy, to act as a direct comparison with [Tayng].

All of the tests were constrained by a bigram/unigram grammer which was built from the F2b corpus. Models were trained using odd-numbered mixtures (Gaussian components) from 1 to 29. Results were obtained for each set of models. Given the enormity of the results (fifteen mixtures for each experiment, multiple weights for some experiments, two baseline experiments, two or more experiments per feature, etc.), only the best results of each set are reported here.

5 Evaluation

The output of the various experiments is evaluated in terms of three basic measures: percent of detected labels which are correct, accuracy (correct - percent of detected labels which are incorrect), and error (100% - 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. Therefore, 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.

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 primary 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 [MAL97], [SBP$^+$92].

6 Results

Tables 2 and 3 show the best results from experiments with non-normalized data, in terms of error. Table 4 show results from experiments with normalized f0 and all thirteen MFCC (all data for all experiments includes first and second derivatives).

It is obvious that zero crossing alone cannot provide any useful input into event detection. The number of insertion errors drove the error to well over 100%. This means that any correct detections were more than cancelled out by insertions. The results of this test are born out when zero crossing data is combined with F0 and energy in unweighted data. The relative error increase of 8% over the baseline result (Base 1) suggested that further experimentation with zero crossing data would be fruitless.

Feature	Alone	With F0 and Energy	Weighted with F0 and Energy	Relative Error to Baseline
Zero Crossings	$>$ 100%	42%	N/A	+8%
Auto-Correlation Peak	74%	39%	37% (0.8 weight)	-4%

Table 2: Error of experiments using zero crossings and auto-correlation peak to augment F0 and energy, with relative error of best result

Auto-correlation peak information was significantly more useful than zero crossing, with an error of 74% when used alone. When added to F0 and energy in unweighted data, the result was a mild relative error reduction of 1%. The relative success of tests on unweighted data encouraged testing on weighted data, which yielded a relative error reduction of 4% against the baseline result (Base 1).

Experiments on using Mel Frequency Cepstral Coefficients in conjunction with non-normalized

F0 data show even greater error reduction than auto-correlation. The initial experiment, in this case, used only the first four coefficients, in order to reduce computing time and space. However, the relative error reduction over Base 1 of 9% encouraged experimentation using all coefficients. The success of weighted data in the auto-correlation peak experiments suggested that weighting would be interesting for these experiments. The result was a relative error reduction of 15% over Base 1. Due to the superiority of this result over the previous results, only MFCC data was used for the tests of normalized data (Base 2).

Instead of simple F0 information for this series of experiments, normalized F0 was used, in order to allow direct comparison of this work and previous work [Tayng]. For this experiment, the HMMs were trained with a weighting of 0.8 for the MFCC data, and tested. Then, a second training and testing cycle was introduced with a weighting of 0.6. The parallel experiments were undertaken in order to determine whether further testing of data weighting were necessary. Both cycles produced improvements over the baseline (Base 2). However, the relative error reduction of the smaller weighting result over the larger one (17%) suggests that data weighting should be tested in further studies.

While the relative error reduction of the MFCC experiments is encouraging, it may 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 an increase in insertions (by improving correct detection). Therefore, an investigation of all three evaluation metrics is useful to determine whether this study has been a success.

Table 5 shows a comparison of the MFCC experiments with the respective baselines and [Tayng]. 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, resulting in a reduction of error. However, one may note that the percentage of insertions has remained the same (17%). This means that the error reduction, while welcome, is not the result of reduced insertions. The results of the normalized data, in contrast, show

Feature	With F0	Weighted with F0	Relative Error to Baseline
Four Cepstral Coefficients	35.5%	36.5% (0.8 weight)	-9%
All Cepstral Coefficients	N/A	32.5% (0.8 weight)	-15%

Table 3: Error of experiments using Mel Frequency Cepstral Coefficients to augment F0, with relative error of best result

Weight	Weighted with F0	Relative Error to Baseline
0.8	37%	-10%
0.6	36%	-12%

Table 4: Error of experiments using Mel Frequency Cepstral Coefficients to augment Normalized F0 and energy, with relative error

	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 5: Comparison of results to baselines and Taylor 1998

both an improvement in correct identification and a reduction of insertions (from 19% to 16%). Thus, 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 is on a wider scale.

7 Conclusion

The large body of experiments that this study encompasses is by no means exhaustive. Future experiments must be done to fill holes in this initial work. A greater testing of the use of data weighting, as well as determining the overall importance of the ngram grammer are tasks both in need of immediate attention. Work on using a tri-gram grammar, different data types, and speaker-independent data are further goals.

In the mean time, it is encouraging that sub-syllable acoustic features have proven useful in improving the overall detection of intonation events. While all future experiments may not be limited to cepstral coefficients, the notable success of their use in this task is promising.

References

[Dus96] K. Dusterhoff. Intone: A prototye intonation analysis system. Master's thesis, Georgetown University, 1996.

[MAL97] 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.

[Mer89] P. Mertens. Automatic recognition of intonation in French and Dutch. In *Proc. Eurospeech 89*, volume 1, pages 46–50, Paris, France, 1989.

[OPSH95] 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.

[SBP$^+$92] 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.

[Tayng] P. Taylor. Analysis and synthesis of intonation using the tilt model. *http://www.cstr.ed.ac.uk/publications/pending/Taylor_pending_d.ps*, forthcoming.

[YJO$^+$96] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *HTK manual*. Entropic, 1996.