USING NEURAL NETWORKS TO LOCATE PITCH ACCENTS

Paul Taylor

Centre for Speech Technology Research, University of Edinburgh, Edinburgh, U.K.
email: Paul.Taylor@ed.ac.uk
http://www.cstr.ed.ac.uk

ABSTRACT

This paper descirbes a technique for finding intonatioanl events, (pitch accents and boundary tones) from waveforms. The technique works in a bottom-up manner by using a recurrent neural network to perform a classification of each frame in the input waveform. An autosegmental description, consisting of intonational events, syllables and the links between them, is then produced from this frame-based classification. The technique correctly identifies 85.7% of pitch accents and boundary tones.

1. INTRODUCTION

In order to use prosodic information in speech recognition, it is necessary to have algorithms which can automatically extract prosodic information from speech waveforms. This paper describes a technique for automatically extracting a representation of an utterance's intonation from its waveform.

The rise/fall/connection (RFC) labelling system [9] achieves good results on a low level intonation labelling task. However, this system falls short of producing the type of output that is ideally required. The most significant problem is that there is no explicit way to link the intonation description to the segmental description. It is possible to say where a pitch accent is located in *time*, but not possible to say which syllable in the utterance is carrying this accent. In addition, the RFC description is too low-level for direct phonological analysis.

This paper describes a completely new technique for extracting intonational information directly from a waveform. The technique is purely bottom-up and requires no additional information, such as the segmental string, to have been pre-calculated.

2. A SYSTEM FOR DERIVING AUTOSEGMENTAL REPRESENTATIONS

The system consists of three basic components, the most important of which is the *intonational event labeller*. The term "intonational event" is a general one used to describe pitch accents and boundary tones. The "event" is the key component in an intonational formalism developed to represent intonation on acoustic, phonetic and phonological levels [10]. The work presented here is only concerned with the finding of events and not with their sub-classification as different sorts of accents and boundary tones. The second component is the *syllable labeller* which produces a list of the syllables in the utterance. The third component is the *linker* which associates events and syllables.

The outputs of the three components combine to form a structure which is in many ways similar to an *autosegmental representation* [3]. In our formalism, this representation consists of two *streams*, each being an ordered list of linguistic units. One stream represents syllables, the other intonational events. *Links* represent associations between units in one stream and units in another[1]. An example is given in figure 1.

An autosegmental representation is particularly useful in intonation, as it allows us to analyse the intonation stream and syllable stream independently. This enables us to compare intonation streams from two utterances directly to see whether they have the same tune or not, while still being able to know if a given syllable is accented or not.

3. THE SYLLABLE LABELLING COMPONENT

Hunt [5] describes a technique for syllable nucleus labelling using recurrent neural networks (RNN). This technique gave 94% correct identification of syllables in a speaker independent test on the TIMIT corpus. Hunt's experiment was replicated on our data with equivalent success (For practical reasons, we used a slightly different RNN topology from Hunt).

3.1. Recurrent Neural Networks

RNNs are now a popular tool in speech and language research as they have a limited ability to model time-dependencies. The type of network used here, com-

[1] This terminology is somewhat different from that of Goldsmith. This is because our autosegmental descriptions are part of a broader linguistic structure formalism which owes its origins to Hertz's [4] Delta system. Our streams are equivalent to Hertz's streams, but it should be noted that our method of relating units on different streams is slightly different to Delta.

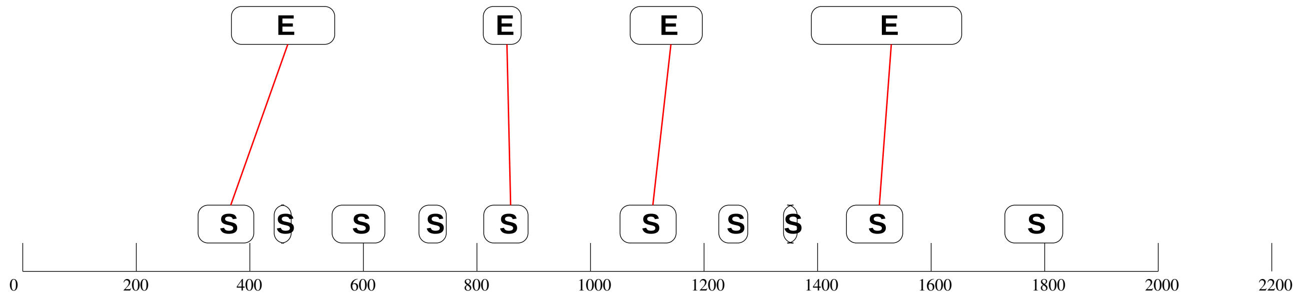

Figure 1. An autosegmental diagram. The top stream shows the position of the intonation events and the bottom stream shows the position of the syllables. The lines indicate the links (associations) between the intonation events and the syllables. The x-axis denotes time and is in ms.

monly referred to as an *Elman* net [2], is similar to a standard 3 layer back-propagation network, but has additional "context" layers which are used for the time dependency modelling. For the syllabification labelling problem, the network consists of an input layer which has the same number of units as each input vector (e.g. one for each cepstral coefficient), a hidden layer of 20 units, and a single output unit. In addition, there are two context layers, one for the hidden layer and one for the output layer. The context layers have the same number of units as their respective hidden and output layers. The network is trained using the back-propagation algorithm.

3.2. Syllabification

Owing to the theoretical difficulties in defining what exactly a syllable is and where its boundaries (if even they exist) should go, syllabification algorithms often simply attempt to delimit syllabic nuclei rather than the syllable boundaries. In nearly all cases syllabic nuclei contain one and only one vowel. The only common exceptions occur due to the presence of syllabic consonants (such as the /m/ in rhythm). Thus an algorithm which can locate all vowels and syllabic consonants can be used for syllabification. This is the approach taken by Hunt, which works as follows.

The technique requires a moderate amount of training data (40 utterances were sufficient) that have previously been hand segmented with phonemic labels.

To train the system, each utterance is analysed in 10ms frame intervals and the RNN is presented with pairs of input and output values in order. There are 12 input values which represent cepstral coefficients and one output value, which is 1.0 if the frame is part of a vowel or syllabic consonant and -1.0 otherwise.

When run, the RNN produces a track against time which indicates how closely a given frame matches a vowel. This vowel-likeness track can then be converted back into a list of labels: every time the track crosses zero in a positive direction a new vowel label is started, and every time it crosses zero in a negative direction the current vowel label is ended, resulting in a series of vowel and non-vowel labels for the utterance. Hunt uses a second RNN to split vowel-vowel sequences into separate syllables (this second stage has not yet been implemented and leads to the syllabification accuracy being worse than the accuracy reported in Hunt's paper).

Hunt's RNN frame classification technique was primarily developed for syllabification, but it can be readily extended to any problem where a binary classification is required to distinguish frames in one broad class from another. This RNN set-up can be used to label vowels (as in syllabification), obstruents, silence etc. This more general technique is referred to as a *binary broad class labeller*.

4. THE EVENT LABELLING COMPONENT

Event labelling can also be formulated as a broad class labelling problem, and so we use the RNN approach for this task also. The output of the net is set up in much the same way: hand marked label files are converted into 10ms interval tracks where a frame has a value of +1 if it lies within the boundaries of an event and -1 elsewhere. In place of cepstral coefficients each frame is associated with a vector of features representing rms energy, smoothed F_0, differentiated F_0, vowel scores, and obstruent scores.

4.1. Smoothed F_0

A "raw" F_0 contour is extracted using a version of the super-resolution pitch tracker, originally developed by Medan et. al [7] and improved by Bagshaw et. al. [1]. Although this algorithm is very accurate in measuring fundamental frequency, further processing is required before the contour can be used.

First of all, there are no F_0 values during unvoiced segments. This is undesirable for intonation recognition as we wish to present contours resulting from the same underlying intonational representations as being the same. The presence of unvoiced segments can superficially make two F_0 contours look different, when in fact they are produced from the same underlying intonational pattern and they are perceived as being equivalent. It is generally accepted [6], [8] that unvoiced segments simply mask the F_0 contour, and therefore the underlying contour can be reconstructed by interpolation through these regions.

A problem arises in that pauses as well as unvoiced segments produce gaps in the F_0 contour. This is problematic as we wish to interpolate through unvoiced regions but not through pauses. To solve this problem,

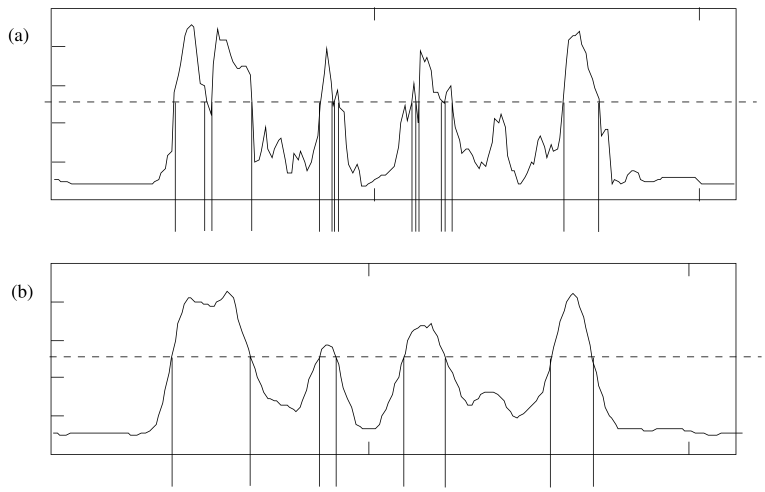

Figure 2. Figure (a) shows the raw output of the event labeller and figure (b) shows the smoothed output. The horizontal dotted line is the zero threshold and the black vertical lines show where the track crosses the threshold. In figure (a) sharp spikes cause the threshold to be crossed several times, resulting in 4 spurious events being inserted. In figure (b), the smoothing has removed all the sharp spikes which results in all the events being labelled correctly.

we use a pause labeller developed using the broad class labelling RNN technique. The two training classes are “non-pause”, containing all speech labels, and “pause” which is mostly silence labels, but also contains some cough and breath noise labels. The pause broad class labeller is very accurate, labelling 99% of frames correctly. Using the pause track, it is possible to fill the gaps in the unvoiced segments while leaving the pauses unfilled.

The F_0 contour is median smoothed using a 170ms long window. The smoothing removes most (but not all) of the perturbations caused by obstruents. The resultant contour is smooth and unbroken in speech regions.

4.2. Differentiated F_0

Regardless of whether one adopts a “tones” or “configurations” view of intonation, it is clear that rapid changes in F_0 values are a a more salient indication of intonational activity than absolute F_0 values. With this in mind a differentiated F_0 contour is calculated from the smoothed F_0 contour, by subtracting the F_0 value of the previous frame from the value of the current frame.

4.3. Vowel scores and Obstruent scores

Previous work [9] showed that substantial rises and falls in the F_0 contour were good indicators of pitch accent presence. It was also shown that other, non-intonational, factors could also produce rises and falls which could potentially lead to confusion if no effort was made to distinguish the two types.

With this in mind, two further tracks were included as input. The vowel track is the same as that used in the syllabification process described in section 3..

Again using the broad class labelling method, an obstruent labeller was developed. This correctly recognised frames within obstruents 98% of the time. This was included as inhibitory information: as obstruents are the primary source for spurious rises and falls, the inclusion of obstruent information can help the RNN identify these spurious F_0 excursions.

4.4. Using the Broad Class Labeller for Event Labelling

The RNN is most sensitive when its input values lie in the range -1 to +1. To facilitate this, the five input variables are normalised. First the mean and standard deviations of each variable are calculated across the entire training set. Next the mean is subtracted from each value which is then divided by twice the standard deviation, so that at least 95% of the values lie in the required range. The RNN is trained as before. When run, the RNN again produces a track with values ranging between +1 (this frame is in an event) and -1 (this frame is not in an event). This track is converted into a label description using the previously described zero-crossing technique.

Experiments have shown that the output of the event labeller is much more uneven than the output of other labellers. The output has a substantial amount of uneven local jitter superimposed on the underlying pattern. This is troublesome when values are close to 0 because fluctuations in this area can cause the zero-crossing technique to insert a large number of spurious labels. To avoid this, the output of the event labeller is smoothed by a 100ms moving average low-pass filter and the resultant track is passed to the zero-crossing algorithm. The raw and smoothed output of the event labelled can be seen in figures 2.

5. THE LINKING COMPONENT

The final component in the system is the linker, which provides the association lines between the syllable and intonation streams. Well formedness criteria govern how the two streams may be linked in principle. Every event must have one and only one link to the syllable stream but a syllable may have zero, one or more links to the event stream. A syllable is normally linked with an event in the case where that syllable is “accented”. Further links

Stream	Events	Syllables
% Correct frames	89.1%	94.5%
Total labels	112	249
Correct labels	98	198
Inserted labels	2	15
% Correct labels	85.7%	79.5%
% Inserted labels	9.1%	6.0%

Table 1. Results showing the accuracy of the event and syllable labellers. (The % inserted labels figure is calculated by dividing the number of insertions by the total number in the original data)

can be made in the case where the syllable is phrase-initial or phrase-final and there are phrase-initial and final intonational events (boundary tones).

At present, only a very simple linking algorithm has been implemented. This algorithm links each event to the syllable whose middle is closest to the event's middle, whilst making sure that all the well-formedness criteria are obeyed.

6. DATA, TESTING AND RESULTS

30 sentences from an American male speaker were used for training and testing. The sentences were part of a natural (but fluent) dialogue. This dataset was labelled by hand with phonemic and intonation labels. The data contained 112 intonation events and 249 syllables. Open-test experiments were carried out in three batches, each time using 20 utterances for training and 10 for testing, and then choosing a different set of training and test utterances for the next batch. In this way, all 30 utterances were subjected to open-class testing.

The labelling algorithms are assessed in two ways. The first measure is a frame by frame measure whereby each frame in the test output is compared to its equivalent frame in the training output. If the training frame's value is -1 and the test value is negative or the training frame's value is +1 and the test value is positive, the frame is judged as correct. This figure is mainly used to assess the raw performance of the neural net.

The second measure determines the accuracy on a label by label basis. Here we calculate how many labels in the training data have been correctly identified and how many spurious labels have been inserted. The definition of "correctly identified" is somewhat arbitrary, but we chose a measure whereby two labels had to overlap by at least 50% to be classed as the same. The results are shown in table 1.

From an analysis of the correctly identified events, the number of links that were correct was 56%.

7. DISCUSSION

The number of correctly identified events, 85.7%, is considerably higher than the 73% accuracy reported for the previous, RFC labeller [9]. In addition, the algorithm is more useful as its output is higher level than the RFC system and also describes the relation between the intonational events and the syllables. Approximately half the errors occurred because two neighbouring events are classified as a single event. More sophisticated processing of the RNN output is required to rectify this problem.

It should be noted however that the linking component of the system does not as yet give a reliable guide as to which event is linked to which syllable. About 40% of the errors occured because the syllable that the event should have been linked to was not recognised. The remainder of the errors were due to failings in the linking algorithm itself. Current work is focusing on building a more sophisticated linker which makes judgements based on the acoustic characteristics of a syllable as well as simply the time distance between it and an event.

ACKNOWLEDGEMENTS

I am grateful to Andrew Hunt who provided his RNN software and gave advice how best to use the RNNs for syllabification purposes. The Stuttgart Neural Network Simulator (SNNS) was also used in the experiments described here and I would also like to thank the writers of that package for their work and for providing this system. As always, Steve Isard provided valuable help and advice on all aspects of the work.

NOTES

All the software and testing programs reported here (written in C++) are available for use by other researchers. The software is in ftp.cstr.ed.ac.uk. Alternatively, consult http://www.cstr.ed.ac.uk.

REFERENCES

[1] P. C. Bagshaw, S. M. Hiller, and M. A. Jack. Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. In *Proc. Eurospeech '93, Berlin*, 1993.

[2] J. L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

[3] John Goldsmith. *Autosegmental and Metrical Phonology*. Blackwell, 1989.

[4] Susan R. Hertz. The delta programming language: an integrated approach to non-linear phonology, phonetics and speech synthesis. In John Kingston and Mary E. Beckman, editors, *Papers in Laboratory Phonology 1*. Cambridge University Press, 1990.

[5] Andrew Hunt. Recurrent neural networks for syllabification. *Speech Communication*, 13:323–332, 1993.

[6] Klaus J. Kohler. A model of German intonation. In Klaus J. Kohler, editor, *Studies in German Intonation*. Universität Kiel, 1991.

[7] Y. Medan, E. Yair, and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Trans. Signal Processing*, 39:40–48, 1991.

[8] Paul A. Taylor. *A Phonetic Model of English Intonation*. PhD thesis, University of Edinburgh, 1992. Published by Indiana University Linguistics Club.

[9] Paul A. Taylor. The rise/fall/connection model of intonation. *Speech Communication*, 15:169–186, 1994.

[10] 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.