

An Automatic Pitch Model with Distance Function

Mohamed Abou-Zleikha, Peter Cahill, Julie Carson-Berndsen

CNGL, School of Computer Science and Informatics, University College Dublin, Dublin, Ireland

mohamed.abou-zleikha@ucdconnect.ie, peter.cahill@ucd.ie, julie.berndsen@ucd.ie

Abstract

Pitch modelling is considered to be an important factor in speech synthesis where the pitch contour plays a demonstrable role in the intelligibility and naturalness of synthesised speech. While quantitative models for pitch contours have been proposed previously, each of these have a fixed level of details and as such not all of them provide the basis either for automatic extraction of pitch model parameters or for measuring the distance between two instances of a model. In this paper, a novel and compact quantitative model for pitch contour is presented which covers the possible variations in pitch and can be automatically extracted. The minimum F0 value, the level global slope of a pitch segment and the semi-periodic jitter properties are used as pitch components and are modelled with a linear function, a sine function and a set of sine functions respectively. A distance measure is defined for the model which takes the shape of the contours into consideration. Experiments show a low mean square error (MSE) for the estimated contours for different languages across different corpora, and investigate the accuracy of the distance function on the model.

Index Terms: pitch modelling, prosody modelling

1. Introduction

While state of the art speech synthesisers can achieve mean opinion scores and word error rates almost on a par with human speakers [1, 2], the difference between current synthesis technology and human speech is more clear-cut when the goal is to produce speech which is not neutral. Excitement, tiredness, anger, and joy are just some of the properties of human speech that current speech technologies cannot convincingly synthesise. Indeed, part of the reason for this is that the extraction and modelling of parameters required to model such intricacies is still an active research topic. It is clear that pitch has an important role in speech [3]. It is one of the components (in addition to, but not limited to: duration, rhythm and energy) that contribute to the additional information in speech that is not in text. Modern pitch estimation algorithms such as the robust algorithm for pitch tracking by Talkin [4], those of Boersma [5] and Yin by de Cheveigné and Kawahara [6], are somewhat competitive in terms of performance, yet how these estimated pitch contours are modelled is still a debated topic.

In concatenative speech synthesis, it is necessary to compare segments of speech to their target specification and to compare segments across concatenation boundaries. Traditionally, such comparisons involve at least acoustic and pitch parameters. Acoustics are often compared by using a distance function (such as weighted Euclidean distance, or Mahalanobis distance). Comparing pitch is less straight-forward, as pitch is only present during voiced segments, where the shape of the contour is more important than the difference between absolute values. The effect of this is that a pitch comparison between an un-

voiced and voiced segment is discrete. Comparing segments which contain continuous pitch contours is often done by subtracting the pitch frequencies (or their log values) at a single point. Such comparisons bear little correlation to the perceived pitch as the local and global trajectories of the pitch are not considered. Calculating delta and delta delta parameters from pitch is not necessarily a solution either as the frames near an unvoiced segment will be represented by the delta of unspecified pitch (i.e. unvoiced) segments.

In this paper, a novel pitch model is presented. The model was designed to overcome the limitations of current approaches, where in particular, it has the following features:

- Automatic application to estimated pitch contours.
- Distance function which measures both global and local properties of the pitch contour.
- Dynamic representation, where the model represents as much variations as the pitch contour requires.
- Speaker- and language-independence.

The remainder of this paper is structured as follows: section 2 reviews existing models, section 3 presents the novel pitch modelling algorithm, section 4 describes evaluations on the model and section 5 concludes and discusses future work.

2. Previous Work

Several pitch models have been proposed over the past 30 years. They were proposed with different intended applications, resulting in models representing different features. Most of them can be classified into the following two categories:

- Symbolic models: where pitch events are represented by a set of symbols.
- Quantitative models: where pitch events are represented by numeric values.

2.1. Symbolic Models

In the symbolic model category, the Pierrehumbert model is one of the first pitch models where the intonation of an utterance is represented as a sequence of high (H) or low (L) tones. H and L are members of a primary phonological opposition [7]. In the IPO model, the intonation contour consists of a linear sequence of discrete intonational elements, movements. Certain movements are perceptually relevant, others are not. Such perceptually relevant changes are represented by positive and negative symbols [8].

The ToBI model is a formulation of tone sequence theory in terms of a transcription system (originally designed specifically for American English but has also been adapted to many languages) [9].

2.2. Quantitative Models

Several quantitative models have been proposed. The quadratic spline model interpolates peaks and valleys of F0 contours with a quadratic spline function [10]. The tilt model generates F0 from tilt parameters which describe the shape of F0 in each intonational event, e.g., pitch accent and boundary tone. The F0 contour of an utterance is represented by a series of these intonational events [11].

The linear alignment model uses curve classes as templates and by combining these curve classes superpositionally it generates F0 contours [12]. The superposition of functional contours (SFC) model simulates intonation by superpositionally combining multiple elementary contours that are functionally defined [13]. The soft-template markup language, based on a soft-template model, describes F0 contours as resulting from realising underlying tonal templates with different amounts of muscle forces under the physical constraint of smoothness [14].

The Fujisaki model represents surface F0 as the logarithmic sum of phrase components and accent or tone components [15]. The target approximation (TA) model is based on the analysis of continuous acoustic data of Mandarin tone and intonation, where it assumes that observed F0 contours are the outcome of implementing pitch targets which are either static (e.g. low) or dynamic (e.g. rise) linear functions [16].

Of the models discussed, where many of them are suggested or used for different tasks like pitch prediction [15, 11], none provide all of the following features: 1) fully automatic modelling from estimated pitch contours, 2) a distance measure that represents the trajectories of the pitch, both globally and locally, 3) Dynamic representation, where the model represents as much variation as the pitch contour requires. Due to the importance of these three features in concatenative speech synthesis, where huge corpora are used, and the importance of the distance function in choosing the units (with respect the features in section 1), a novel model is presented in next section.

3. The Pitch Model

The model presented in this paper is quantitative, where the pitch contour is represented as a sum of functions. The number of functions used controls the precision of the model. Each function corresponds to a level of granularity of the model, where the highest level is the *command level*. The command level models the global pitch of a continuously voiced segment of speech. In practice this is often a word, but it may also be part of a word, or a combination of voiced words. The other levels in the model represent jitter, where each additional level represents a progressively finer grained level of semi-periodic jitter.

Several studies have been carried out on global slope with respect to pitch variation, where it has been shown that the phrase pattern in pitch contours resembles a Gaussian function shape as in the Fujisaki model [15]. Thus any pitch segment can be represented by a Gaussian or part of Gaussian function. This is true when the segment contains only one phrase command, but if the segment contains more than one phrase command, the Gaussian function would estimate the first one and fail to estimate the others, and in that case more than one Gaussian function is needed to estimate the segment. For this reason, the command level is represented by a sine wave, where each oscillation corresponds to a phrase command. While the timing of the sine wave correlates well to phrase commands, it is not necessarily true that the amplitudes of the sine wave will match

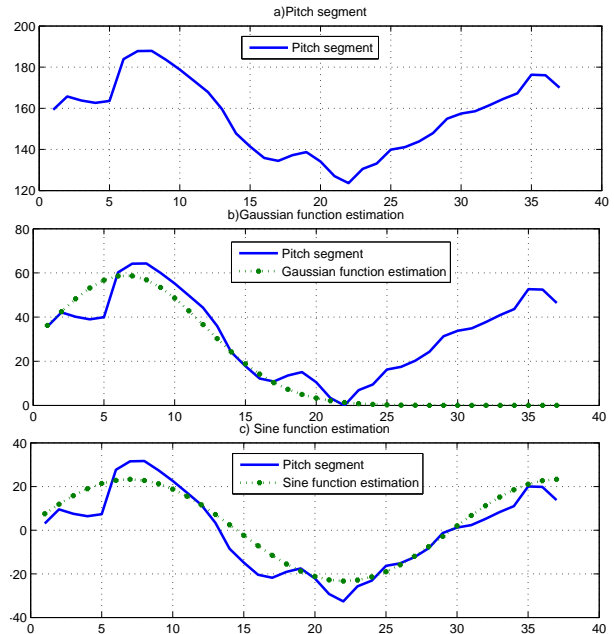


Figure 1: Comparing the estimation between the sine function and the Gaussian function for the command level estimation of a pitch segment, parts: (a) represents a voiced segment with 2 phrase commands, (b) Gaussian estimation and (c) sine estimation.

all phrase commands. This difference is compensated by the *first jitter level*. Figure 1 illustrates the difference in estimation between the sine function and the Gaussian function for a pitch segment, where it is clear that the Gaussian function failed to estimate the second phrase command of the segment, while the sine function estimated it correctly.

The model assumes that all jitter levels are semi-periodic. This is similar to other work [17], where non-periodic jitter, while important, is considered random and is thus not possible to model. Therefore, the command and periodic jitter levels can be represented as a sum of periodic functions.

While any periodic function is appropriate for the model, in the case of this work, the focus is on sine functions. The sine function can be defined as

$$S(x) = a \sin(bx + c), \quad (1)$$

where, the amplitude, a , is the peak deviation from the centre position. The frequency, b , specifies how many oscillations occur in a unit time interval, and the phase, c , specifies where in the cycle the sine wave starts. A shifting parameter is subtracted from the pitch segment before the estimation of each sine function parameter. Figure 2 illustrates the parameter estimation process, where each sine wave is extracted in an iterative process that minimises the error. This method is similar to harmonic decomposition [18].

As a result, the pitch segment representation contains the following parameters:

- min F0,
- begin time,
- end time,

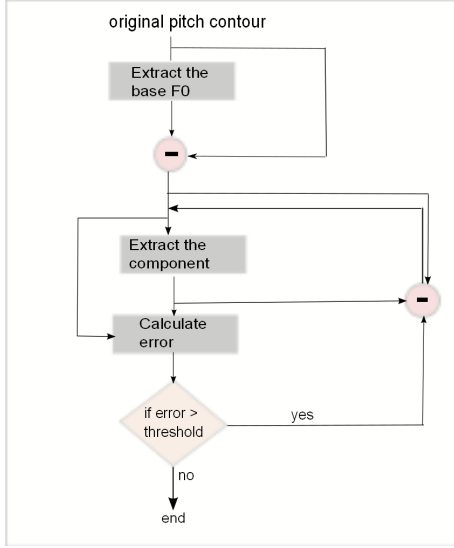


Figure 2: Pitch parameter estimation process

- a command level component function, which consists of:
 - pa : amplitude of the component,
 - pb : frequency of the component,
 - pc : phase of the component,
 - $pshift$: shifting parameter,
- a set of jitter sine functions, each consisting of:
 - a : amplitude of the jitter,
 - b : frequency of the jitter,
 - c : phase of the jitter,
 - $shift$: shifting parameter.

As the command level component and jitter level components are all sine functions, they can be merged into one estimation formula as follows:

$$F0(t) = F0min + (pa \sin(pb * t + pc) + pshift) + \sum_{i=1}^n (a_i \sin(b_i * t + c_i) + shift_i) \quad (2)$$

where n is the number of sine functions used to estimate jitter in the contour. The experiments showed that two sine functions can achieve a reasonable approximation of the pitch contour, although more precision can be obtained by using more jitter functions.

Figure 3 illustrates the application of the model to a pitch contour, where the command level and two jitter levels are modelled. The original pitch contour is illustrated in figure 3a and figures 3b- 3f represent the steps of the estimation process.

Step 1: Identify the minimum pitch value during the contour, where in the case of this example, the value is approximately 80Hz, as illustrated in figure 3b.

Step 2: The minimum pitch value is then subtracted from all points in the pitch contour and the contour is then fitted to the first sine function, representing the command level as in figure 3c.

Step 3: After the fitting process is complete, the fitted sine wave is subtracted from the previous step, and the result is then

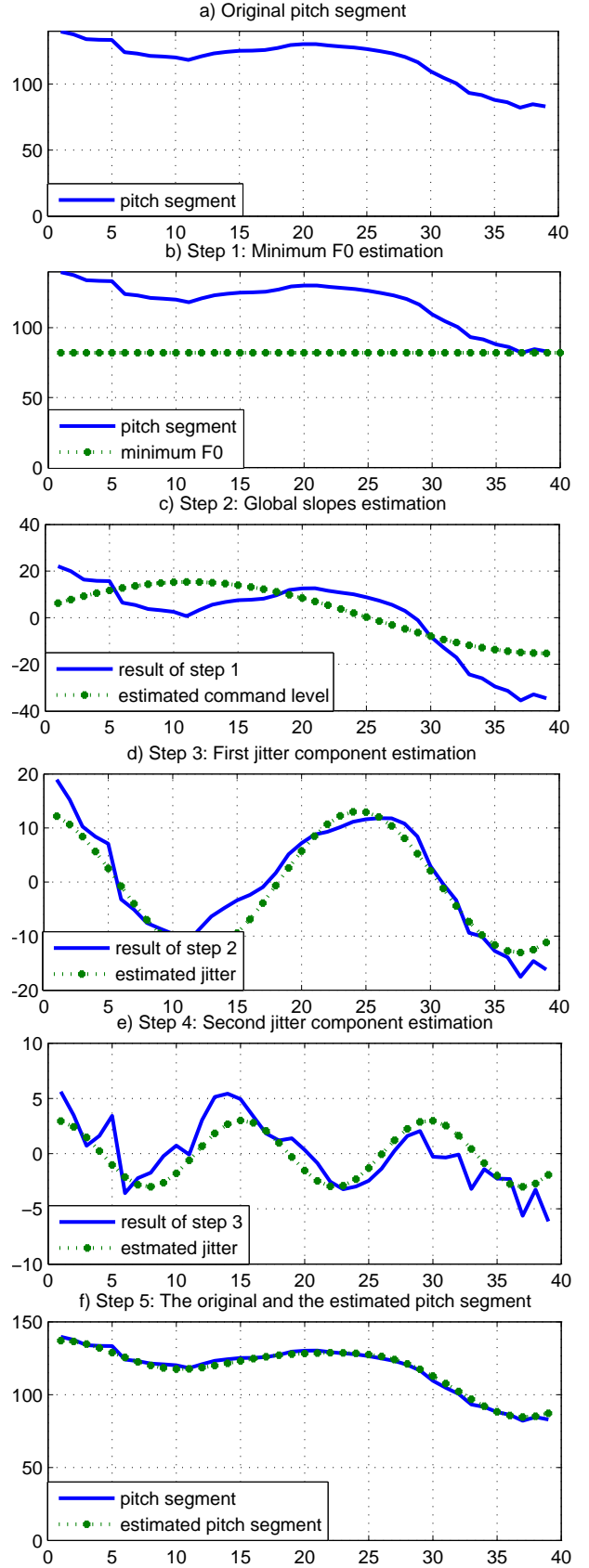


Figure 3: Estimation process of a pitch segment.

fitted to another sine wave, This represents the first jitter level, as in figure 3d.

Step 4: This process is repeated again to represent the second jitter level, as illustrated in figure 3e.

Step 5: By combining the minimum pitch value with the 3 sine functions as in equation 2, the pitch contour is then modelled as in figure 3.

3.1. Distance Measurement

Comparing two instances of a pitch model is not a straightforward process, since the similarity between the shapes of the contours is more perceptually relevant than the distance between absolute values. This approach suggests that three factors need to be considered in calculating the distance:

1. The duration of the segments.
2. The minimum value of $F0$ over the segments.
3. The function parameters.

When comparing segments with respect to $F0$, it is quite likely that the duration of each segment is different. Similarly, the minimum respective values of $F0$ are likely to differ. However, $F0$ contour segments that have slightly different minimum $F0$ but are very similar with respect to command level and semi-periodic jitter are likely to be perceptually much more similar than segments which have matching minimum $F0$, but with different trajectories.

3.1.1. Duration Scaling

In order to ensure the duration of the segments are equal, the shorter of the segments is scaled to match the duration of the longer segment. Given two sine functions

$$S1(a1, b1, c1, shift1) \text{ defined on } 1..w$$

$$S2(a2, b2, c2, shift2) \text{ defined on } 1..v$$

where $v < w$, the range of $S2$ needs to be modified to make it equal to the range of $S1$:

$$p = (w - 1)/(v - 1)$$

$$Newa2 = a2 * p$$

$$Newb2 = b2/p$$

$$Newa2 = c2$$

$$Newshift2 = shift2$$

After this process both $S1$ and $S2$ have equal duration, the distance between the two segments can be calculated from all of the parameters involved in the model. In order to do this, measuring the distance between two sine functions is considered first.

3.1.2. Sine Function Distance

Given the following two sine functions that have the equal duration

$$S1(a1, b1, c1, shift1)$$

$$S2(a2, b2, c2, shift2)$$

The parameters of $S1$ and $S2$ can be used to calculate a distance between the two pitch segments. In order to make the distance function perceptually relevant, it is assumed that the shape of a pitch contour is more important than the amplitudes

of the contour. The distance function therefore gives a greater weight to the frequency and phase parameters. Rather than using the absolute difference in amplitudes, the natural logarithm of the difference is used.

$$if(|a1 - a2| > 1)$$

$$difA = \log|a1 - a2|$$

$$else$$

$$difA = 0 \quad (3)$$

The distance between them can be defined as follows:

$$dis(S1, S2) = \alpha difA + \beta |b1 - b2| + \gamma |c1 - c2| + \theta |shift1 - shift2| \quad (4)$$

where $\alpha, \beta, \gamma, \theta$ are the weights of the amplitude, frequency, phase, and shift respectively.

By adding the duration scaling effect, where the difference between the duration of the functions is proportional to the distance between them:

$$dis \propto \delta \text{ where } \delta = |w - v|$$

The distance function becomes:

$$Sdis(S1, S2) = dis(S1, S2) + \lambda * \delta \quad (5)$$

where λ is the weight of duration scaling.

The final distance function is a combination of the described functions as follows:

$$finalDis = \sum_{i=1}^r (\mu_i Sdis(S1_i, S2_i)) + \xi |minF0_1 - minF0_2| \quad (6)$$

where μ_1 is the weight of the command level component, $r - 1$ is the required number of sine functions to estimate the jitter components, $\mu_i : i = 2..r$ are the weights of the jitter components, and ξ is the weight of the difference between the minimum pitch values.

4. Results

To evaluate the model, three experiments have been undertaken. The first was to identify how many levels of jitter should be extracted to model prosodically rich speech. The second experiment calculated the error of the modelling algorithm; and the third investigates the accuracy of the distance function.

4.1. Identification of Jitter Levels

For the first experiment, a 950 utterance subset of the single speaker, received pronunciation British English database "Roger" by the Centre for Speech Technology Research, University of Edinburgh, was used [19]. The subset of the corpus used for the experiment was prosodically rich parts of Lewis Carroll's children stories.

Figure 4 shows the mean square error (MSE) of the model when using different amounts of jitter sine functions, where MSE is the error of the pitch estimation from the model. The more error the model has, the more detail is lost in the estimation. For the Roger corpus, Two sine functions for jitter is enough where it is clear that the difference in having more than two jitter sine functions is comparatively small. For alternative applications of the pitch model or corpora, a different number of jitter sine functions may be appropriate.

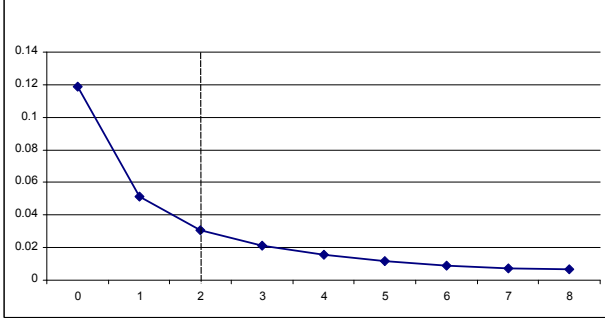


Figure 4: MSE per number of jitter sine functions.

4.2. Robustness Across Speakers and Languages

For the second experiment, 34687 sentences from the Oregon Graduate Institute 21 Languages Corpus [20] were used. Each language is represented by 1700-2500 utterances. The corpus consists of telephone-quality speech with multi-speaker responses for questions.

For two jitter components, MSE for each language has been calculated, and the results are shown in the table 1.

The model achieved similar results across all languages, including tonal languages. These results are also consistent with the high quality Roger voice data as used in the first experiment.

Table 1: MSE per language.

Language	Number of sentences	MSE
Arabic	2035	0.0183
Cantonese	2510	0.0157
Czech	2197	0.0111
American English	2394	0.0118
Farsi	1787	0.0093
German	2475	0.0125
Hindi	2473	0.0133
Hungarian	2494	0.0098
Indonesian	2506	0.0144
Italian	2211	0.0170
Japanese	2351	0.0135
Korean	2521	0.0105
Mandarin	2461	0.0119
Polish	2494	0.0096
Portuguese	2508	0.0120
Russian	2474	0.0131
Spanish	2515	0.0128
Swahili	1718	0.0172
Swedish	2526	0.0158
Tamil	2401	0.0176
Vietnamese	2345	0.0211
sum:	49396	avg: 0.0137

The prosodically rich "Roger" corpus has a higher MSE than the 21 languages corpus. This is because the 21 languages corpus contains less prosodic variations, where more variations in pitch require more jitter levels in the pitch model. This verified the importance of a dynamic number of jitter levels in the model.

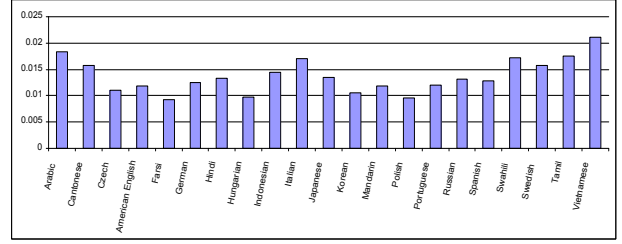


Figure 5: MSE per language

4.3. Accuracy of the Distance Measure

For the third test, a subset of the Roger corpus was used for the experiment. The selected subset contains 1193 words, where each word is spoken in one of the following prosodic classes: question, declarative, or exclamatory. Two classification algorithms were used: feed forward back propagation neural networks and hierarchical k-medoids clustering.

4.3.1. Neural Networks

The model that was discussed in the section 3 has been used as an input to train a neural network. The input layer consists of 15 features (the parameters of the model for two jitter components), several configurations have been tested, where the best results come from a neural network with 1 hidden layer and 70 neurons. The output is a single neuron, where the value is rounded between 1 and 3 (representing the class index). The training algorithm is the gradient descent backpropagation with adaptive learning rate.

4.3.2. Hierarchical K-Medoids

K-medoids clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centre, the centre is defined as the data point in each cluster which is closest to the mean of the cluster. As a result, n clusters are generated and to classify a new example, the distance between the features of the example and the centre of each cluster is calculated, where the cluster with the minimum distance to the example is selected as the class of the example.

In the hierarchical k-medoids algorithm, the data of each cluster is recursively divided using the k-medoids algorithm according to a criterion which is explained below. A semi supervised learning version of hierarchical k-medoids was used, where the data is labelled and these labels are used to calculate the number of occurrences (count) of each class. If there are x classes that achieve the condition $count(class) > threshold$, the cluster data are re-clustered to x clusters and the process is repeated for the remaining clusters. The used weights of the distance function is shown in tables 2,3.

4.3.3. The Results

The results of the experiment show that the hierarchical k-medoids method achieved better results when the number of clusters is high. It achieved 90.6% recognition rate versus 86.3% for the neural network, but this recognition rate decreases when the number of clusters is decreased by increasing the majority threshold. The recognition rate of the system is shown in table 4.

Table 2: *The weights of distance function components.*

Component	Weight
Command Level	1
First Jitter Level	1
Second Jitter Level	0.5
Min F0	0.01

Table 3: *The weights of sine function components.*

Component	Weight
a	0.5
b	1
c	1
shift	1
scaling	0.1

5. Conclusion

This paper presents a novel pitch model, which was motivated by the need for a model that can be automatically applied to estimated pitch contours, contain a perceptually relevant distance function and a dynamic precision, where the model represents as much variation as the pitch contour requires. While the primary motivation for this work is to improve the F_0 modelling for concatenative speech synthesis, it is not limited to this task.

Several experiments were performed on the model. The first experiment evaluates the mean square error with respect to the number of jitter modelling sine functions. From the experiment, it is clear that two jitter sine functions are appropriate, although moderately better performance can be achieved by using more. The second experiment verified that the models performance is consistent, both with different speakers and different languages. The third experiment evaluates the accuracy of the distance measure.

The results of the experiments are promising. However, more tests are needed using a unit selection speech synthesiser, where the model and the distance function will be part of the prosody model of the synthesiser. Future work will involve evaluating the distance measure of the model in perceptual tests, by integrating the model into a unit selection speech synthesiser.

6. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin, Ireland. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland. The authors would like to thank the Centre for Speech Technology Research, University of Edinburgh, for access to the Roger corpus.

7. References

- [1] V. Karaiskos, S. King, R. Clark, and C. Mayo, "The blizzard challenge 2008," *Proceedings of Blizzard Challenge Workshop*, 2008.
- [2] S. King and V. Karaiskos, "The blizzard challenge 2009," *Proceedings of Blizzard Challenge Workshop*, 2009.
- [3] W. Hess *et al.*, "Pitch determination of speech signals: algorithms and devices," 1983.

Table 4: *Recognition rate of the classification algorithms.*

Algorithm	Recognition rate
Neural Networks	86.3%
hierarchical k-medoids (44 cluster)	90.6%
hierarchical k-medoids (16 cluster)	88.3%
hierarchical k-medoids (3 cluster)	82.2%

- [4] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.
- [5] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, vol. 17, 1993, pp. 97–110.
- [6] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, p. 1917, 2002.
- [7] J. Pierrehumbert, *The phonology and phonetics of English intonation*. MIT Cambridge, MA, 1980.
- [8] R. Collier and A. Cohen, *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge University Press, 1990.
- [9] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," in *Second International Conference on Spoken Language Processing*, vol. 2, 1992, pp. 867–870.
- [10] D. Hirst, "Automatic modelling of fundamental frequency using a quadratic spline function," *Travaux de l'Institut de Phonétique d'Aix*, vol. 15, pp. 71–85.
- [11] P. Taylor, "The tilt intonation model," in *Fifth International Conference on Spoken Language Processing*, 1998.
- [12] J. van Santen and B. Möbius, "A model of fundamental frequency contour alignment," *Intonation: Analysis, Modelling and Technology*, 1999.
- [13] G. Bailly and B. Holm, "Sfc: a trainable prosodic model," *Speech Communication*, vol. 46, no. 3–4, pp. 348–364, 2005.
- [14] G. Kochanski, C. Shih, and H. Jing, "Quantitative measurement of prosodic strength in mandarin," *Speech Communication*, vol. 41, no. 4, pp. 625–646, 2003.
- [15] H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech," in *Speech Prosody 2004 International Conference*. ISCA, 2004.
- [16] C. Xu, Y. Xu, and L. Luo, "A pitch target approximation model for f0 contours in mandarin," in *Proceedings of The 14th International Congress of Phonetic Sciences*, 1999, pp. 2359–2362.
- [17] J. Schoentgen and R. De Guchteneere, "Predictable and random components of jitter," *Speech Communication*, vol. 21, no. 4, pp. 255–272, 1997.
- [18] E. George and M. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 389–406, 1997.
- [19] V. Strom, R. Clark, and S. King, "Expressive prosody for unit-selection speech synthesis," in *Ninth International Conference on Spoken Language Processing*, 2006.
- [20] T. Lander, R. Cole, B. Oshika, and M. Noel, "The ogi 22 language telephone speech corpus," in *Fourth European Conference on Speech Communication and Technology*. ISCA, 1995.