Explicit Duration Modelling in HMM-based Speech Synthesis using a Hybrid Hidden Markov Model-Multilayer Perceptron

Kalu U. Ogbureke, João P. Cabral, Julie Carson-Berndsen

CNGL, School of Computer Science and Informatics, University College Dublin,
Belfield, Dublin 4, Ireland

kalu@ucdconnect.ie, joao.cabral@ucd.ie, julie.berndsen@ucd.ie

Abstract

In HMM-based speech synthesis, it is important to correctly model duration because it has a significant effect on the perceptual quality of speech, such as rhythm. For this reason, hidden semi-Markov model (HSMM) is commonly used to explicitly model duration instead of using the implicit state duration model of HMM through its transition probabilities. The cost of using HSMM to improve duration modelling is the increase in computational complexity of the parameter re-estimation algorithms and duration clustering using contextual features. This paper proposes to use an alternative explicit duration modelling approach to HSMM which is a hybrid of HMM and multilayer perceptron (MLP). The HMM is initially used for state-level phone alignment, in order to obtain state durations of HMM for each phone. In the second stage, duration modelling is done using an MLP where the inputs are contextual features and the output units are the state durations. Both objective and perceptual evaluations showed that the proposed duration modelling method improved the prediction of duration and the perceptual quality of synthetic speech as compared with HSMM.

Index Terms: duration modelling, HMM-based TTS, hidden Markov model, multilayer perceptron

1. Introduction

HMM-based speech synthesis is the parametric method that produces the highest quality and offers great parametric flexibility for transforming voice characteristics, e.g. by using adaptation techniques [1, 2]. Duration is an important aspect of speech related to prosody, which has a great effect on the perceptual quality and expressiveness of synthetic speech. Furthermore, in some languages like Finish, the duration of phonemes conveys meaning, e.g short and long phonemes convey different meanings. Thus, errors in duration prediction can change the meaning of a word [3].

The duration of speech can be implicitly modelled by the transition probabilities between HMM states. However, the distribution that results from this implicit modelling is exponential which is not appropriate for modelling the duration of phones as the duration of phones are generally normally distributed [4]. In order to overcome this problem in statistical speech synthesis, duration is explicitly modelled by using HSMM [5]. In this method, state duration is usually modelled using single Gaussian distributions and the duration models are clustered and tied using decision trees to deal with data scarcity as well as the problem of estimation of the duration of phone contexts not seen during training. Since HSMM is used as a generative model in speech synthesis, the duration of synthetic speech is represented by the number of speech frames generated from each state, based on the state duration distributions.

The motivation of this work is the improvement of the accuracy of duration prediction in HMM-based speech synthesis in order to improve the perceptual quality of synthetic speech. The approach presented is a development of a previous work in [6] where duration is explicitly modelled using continuous HMM. In the previous work, a decision tree is used to predict the durations of models not seen during training while in the present work, an MLP is used. The proposed duration modelling approach is a combination of HMM and MLP. HMMs are used in a first stage to obtain initial estimates of phone durations. In this process, monophone HMMs are trained using parameters of the speech spectrum, followed by state-level alignment of the training data. State durations are estimated from the alignment as the number of observations assigned to each HMM state. In a second stage, duration is modelled by training an MLP using phonetic, prosodic and articulatory features from context-dependent phone labels. The output units of the MLP represent the state durations (number of frames) obtained in the first stage.

The next section gives an overview of the baseline explicit duration modelling method using HSMM. Section 3 describes the proposed method for explicit duration modelling using a hybrid HMM-MLP method. In Section 4, this proposed approach is compared with the baseline method in terms of both an objective and a subjective evaluation. Finally, conclusions are presented in Section 5.

2. Explicit duration modelling using HSMM

In HSMM, the state duration is modelled with an n stream Gaussian distribution, where n represents the number of states in the HSMM. Figure 1 shows a HSMM with 5 states. In HSMM, $p_j(d)$ (the state duration distribution for state j) is explicitly modelled with a Gaussian distribution. d is the duration of each state and $b_j(o)$ is the state emission probability for state j, while o is the observation. During training, the duration and model parameters are re-estimated using algorithms such as Baum-Welch.

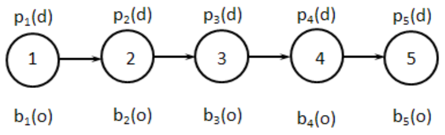

Figure 1: A 5-state HSMM with explicit duration density represented by $p_j(d)$.

Due to the large number of context-dependent factors, that are necessary to take into account in modelling duration, the parameters of the duration distributions might not be robustly estimated for models with small number of occurrences in the training corpus. Furthermore, duration models for phone contexts not seen in the training corpus need to be estimated from context-dependent models obtained during training. These problems are usually addressed using decision trees. In order to cluster and tie the parameters of state distributions, the duration distributions of all streams are entered at the root node of a tree. Then phonetic and prosodic context questions are asked at each node and depending on the answers, the states are split using minimum descriptive length (MDL) criterion [7]. The split operation continues until all questions are asked. In the lower part of Figure 2, the leaf nodes contains four clusters (A-D) whereby each cluster is tied. Clusters that are tied share common parameters of the duration distributions, namely, the means and variances. For the estimation of the duration distributions of contexts not seen during training at the synthesis stage, the decision tree is traversed from the root to the leaf node.

The upper part of Figure 2 shows an example of a phone model and the respective context-dependent label comprising of phonetic and prosodic features (represented by symbols after the symbol '@') and the lower part shows clustering and tying of this model using decision tree. In this example, the phonetic context question 'C-Central_Fricative' asks if the current phone belongs to central fricative class and the prosodic context questions 'L-Syl_Stress' and 'C-Syl_Stress' deal with stress on the previous and current syllable, respectively. For example, in the HTS speech synthesis (version 2.1) system for English demo [8], each phone has 53 phonetic and prosodic features which deal with phone identity, syllable, words, parts-of-speech, phrase and utterance information.

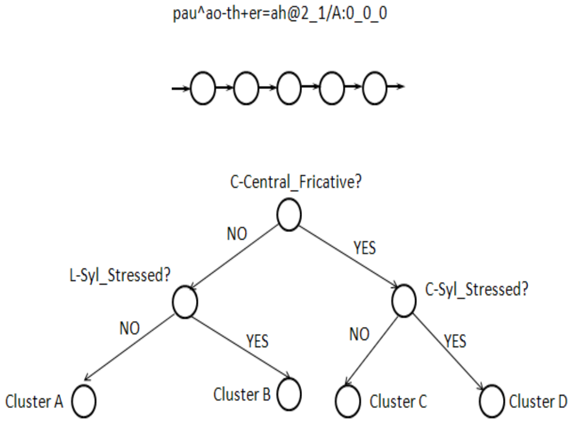

Figure 2: Illustration of decision tree-based clustering for context-dependent label of HSMM (top) and duration modelling (bottom).

3. Duration modelling using a hybrid HMM-MLP

This section describes the proposed explicit duration modelling method using a hybrid HMM-MLP. MLP has been previously used to model segmental durations in speech synthesis, e.g. [9, 10]. The approach presented in this paper is different from these works in that it is applied to HMM-based speech synthesis. There are two training stages involved, namely, the training of the alignment model followed by explicit duration modelling as shown in Figure 3. These two parts are described in the following sections.

3.1. Alignment model

Phonetic alignment is the process of finding the phone boundaries for a speech segment given the phone sequence for that segment. The technique commonly used for automatic phonetic alignment is the Viterbi algorithm. It determines the best state sequence, given a phone sequence and a sequence of speech frames. In this work, the duration of each state which is given in number of frames is obtained by dividing the duration, in milliseconds, by the frame rate.

3.2. Explicit duration modelling using MLP

3.2.1. MLP architecture

MLP is made of simple processing units which communicate by sending signals to each other over a large num-

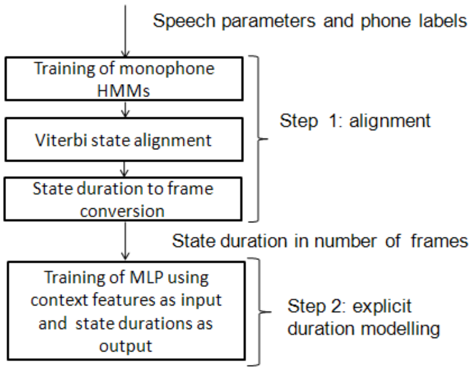
Speech parameters and phone labels
Training of monophone HMMs
Viterbi state alignment
State duration to frame conversion
Step 1: alignment
State duration in number of frames
Training of MLP using context features as input and state durations as output
Step 2: explicit duration modelling

Figure 3: Training stages of explicit duration model using a hybrid HMM-MLP.

ber of weighted connections [11]. MLP has at least two layers of processing units. The most common in speech processing is a two-layer perceptron which has an input layer with non-processing units and both a hidden and output layers with processing units. The output layer processes the signal propagated from the input layer, through the hidden layer, and outputs the result which may be further processed depending on the application (e.g. scaling, conversion, etc.).

The MLP architecture used in this work to predict state durations is shown in Figure 4. The five units in the output layers (dur_s1, dur_s2, dur_s3, dur_s4 and dur_s5), represent the state durations for states 1 to 5 (for HMMs with 5 states) respectively, obtained from the alignment stage. The *tanh* and *linear* activation functions are used in the hidden and output units respectively. The activation function scales the state durations to be within a given range. The *tanh* function scales the input to be between -1 and $+1$ and the scaling factor of the *linear* function was set equal to 1 (the input and output values are the same). The input features $F1 - F128$ represent the phonetic, prosodic and articulatory features extracted from each phone. These features are described in the next section

3.2.2. *Phone context features*

The phone features used as input of MLP comprise the original 53 set of features used by the baseline speech synthesis system described in Section 2 plus a set of 25 articulatory features. The latter set of features is used for the previous, current and next phone, giving a total of 128 features for each phone. Symbolic features like parts-of-speech and phone identity are represented with distinct numerical values. For example, parts-of-speech feature with symbolic values $\{aux, content, det, pps\}$ are represented by $\{1, 2, 3, 4\}$. The set of articulatory features of Table 1 was originally used in [12] for speech recognition and is used in this work. The articulatory features are binary, for example, if a phone is a fricative, the value is 1, otherwise the value is 0.

Articulatory features
approximant, fricative, glottal, nasal retroflex, stop, vocalic, voiced
alveolar, dental, labial, palatal palveolar, velar
back, central, front
high, low, mid, semihi, semilo
round, static, tense

Table 1: List of articulatory features used as input features of the MLP.

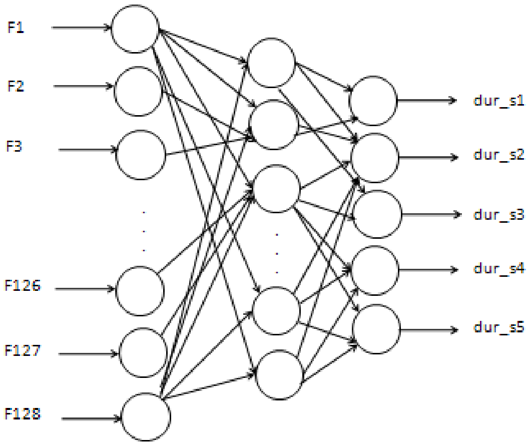

Figure 4: Architecture for training MLP to predict state durations.

3.2.3. *Speaking rate control*

Similarly to HSMM, the speaking rate can also be controlled in MLP. This is done by scaling the MLP weights. There are two sets of weight matrices, namely, W_1 of dimension n_{in} x n_{hid} and W_2 of dimension n_{hid} x n_{out}. Where n_{in} is the number of input features to the MLP, n_{hid} is the number of hidden units and n_{out} is the number of units in the output layer. The speaking rate can be controlled as follows:

$$\hat{W}1 = \beta W1, \tag{1}$$

$$\hat{W}2 = \beta W2, \tag{2}$$

where $\hat{W}_1$ and $\hat{W}_2$ are the transformed weight matrices and β is a positive scaling factor. Fast rate is achieved when β is less than 1 and slow rate is achieved when β is greater than 1.

Figure 5 illustrates the prediction of duration using a two-layer MLP with one unit in the hidden and output layers. x represents the input duration while $w_1 \in W_1$ represents the weight from the input to the hidden layer and $w_2 \in W_2$ represents the weight from the hidden to the output layer. The activation functions used in the hidden and output layers are $f_1(a)$ (*tanh*) and $f_2(a)$ (which is *linear*) respectively while a represents the activation. The predicted duration y is determined as follows:

$$y = w_2 \left(\frac{e^{w_1 x} - e^{-w_1 x}}{e^{w_1 x} + e^{-w_1 x}} \right), \qquad (3)$$

while the speaking rate can be controlled, during synthesis, as follows:

$$\hat{y} = \hat{w_2} \left(\frac{e^{\hat{w_1} x} - e^{-\hat{w_1} x}}{e^{\hat{w_1} x} + e^{-\hat{w_1} x}} \right), \qquad (4)$$

$$\hat{y} = \beta w_2 \left(\frac{e^{\beta w_1 x} - e^{-\beta w_1 x}}{e^{\beta w_1 x} + e^{-\beta w_1 x}} \right). \qquad (5)$$

This is illustrated in Figure 3.2.3 for $x = 2$, $w_1 = 3$ and $w_2 = 2$. The evaluation of speaking rate control is beyond the scope of this work.

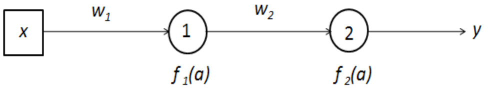

Figure 5: An illustration of duration prediction in MLP.

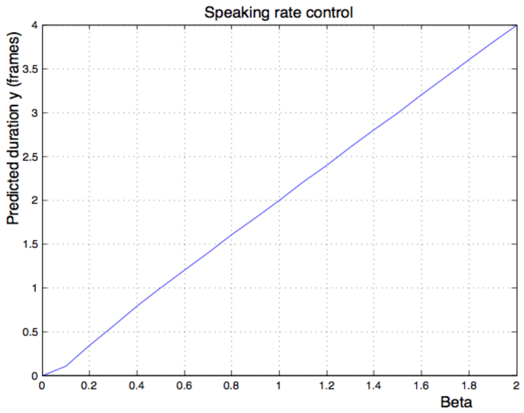

Figure 6: An illustration of speaking rate control in MLP.

4. Evaluation of duration modelling in HMM-based speech synthesis

Three versions of HMM-based speech synthesisers are used in these experiments which differ in the method for modelling duration. The baseline system uses the HSMM as described in Section 2, whereas the second system uses the proposed HMM-MLP approach presented in Section 3. The third system is the baseline system which uses the natural durations of the speech.

4.1. Speech corpus

The RMS voice of CMU_ARCTIC corpus [13] of read speech was used for training of acoustic models as well as testing. The corpus was divided into a training, test and development set composed of 1030, 82 and 20 sentences respectively. The development set was used to choose the optimum number of hidden units of the MLP.

4.2. HSMM-based speech synthesiser

4.2.1. Analysis

The F0 parameter was estimated using the implementation of the RAPT algorithm [14] of the Entropic Speech Tools (ESPS). Besides the F0, the spectral envelope of the speech signal and the aperiodicity spectrum for each frame was estimated using the STRAIGHT method [15].

4.2.2. Statistical modelling

The statistical modelling and parameter generation were implemented using the HTS toolkit version 2.1 [16]. The parameters used were the 24th order mel-cepstrum, F0 and five aperiodicity parameters, with their delta and delta-delta features. HSMM with three streams were used for statistical modelling of the F0, aperiodicity and spectrum parameters respectively. During HSMM training, each stream for spectrum, F0 and aperiodicity was clustered using different decision trees to deal with data sparsity as well as to predict unseen contexts. The number of leaf nodes of the decision tree for duration was 492.

4.2.3. Synthesis

During synthesis, speech parameters were generated by the HSMMs from the sentences in the test set and then the speech waveform was generated from the parameters using the STRAIGHT vocoder. Speech was also synthesised from the generated parameters but imposing the durations from the proposed approach and durations measured on recorded speech.

4.3. Duration modelling using HMM-MLP

The number of units in the hidden layer of the MLP was determined experimentally on the development set. Figure 7 shows the variation of the Root Mean Squared Error (RMSE) of phone duration averaged over all phones, in milliseconds, relatively to the number of units in the hidden layer on the development set. RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \hat{x}^2}, \qquad (6)$$

where $\hat{n}$ is the number of occurrences of each phone in the train, test or development set respectively and $\hat{x}$ is the difference between the reference and predicted durations.

The reference durations were obtained from the phone annotations. The optimal number of units in the hidden state was 75.

An MLP with 128, 75 and 5 units in the input, hidden and output layer respectively, was trained using an implementation of the *backpropagation* algorithm [17].

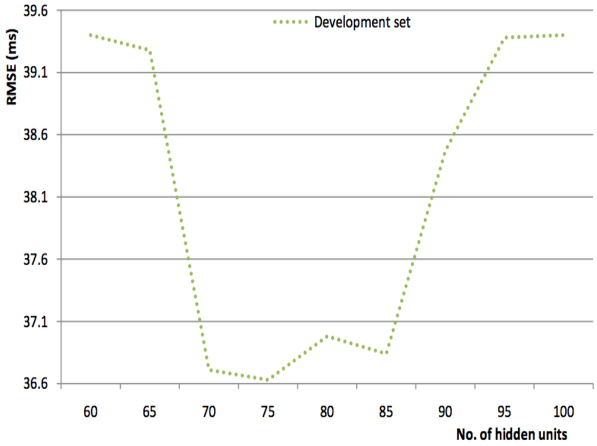

Figure 7: Variation of mean value of RMSE for all phones as a function of the the number of units in the hidden layer.

4.4. Objective evaluation

4.4.1. Measurement

The criterion used for the objective evaluation is the RMSE between the predicted and reference (measured on recorded speech) phone durations given by (6).

4.4.2. Results

Figure 8 shows the RMSE in milliseconds (ms) obtained for the phones in the test set. Figure 9 shows the mean RMSE for all the phones in the train, test and development sets respectively. The latter shows that the proposed system obtained lower mean RMSE for the train, test and development sets.

Also, some phones are better modelled with MLP, e.g. the phones ‘ey’, ‘uh’ and ‘uw’. The phones with the best and worst performance are ‘ax’ and ‘ng’ respectively in both systems. Furthermore, the proposed system performed better than the baseline in most of the phones, while the proposed system poorly performed on the ‘aw’ phone.

4.5. Subjective evaluation

The effect of using the proposed duration model on the perceptual quality of synthetic speech was evaluated by conducting an ABX forced-choice test.

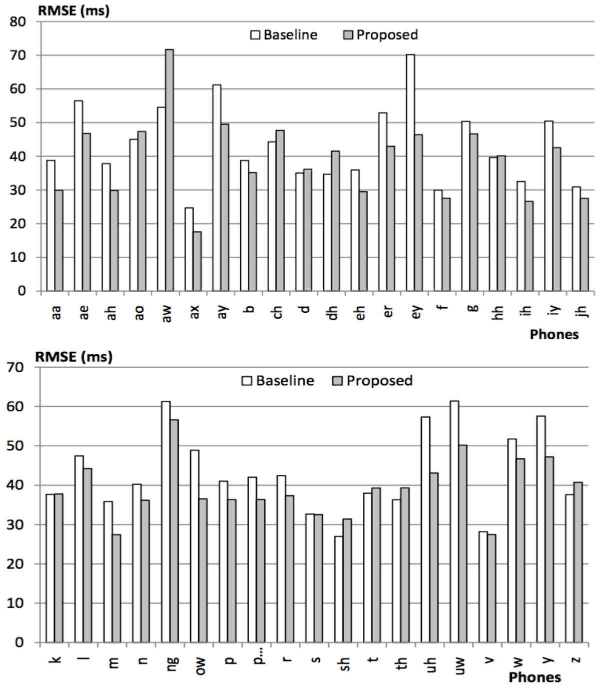

Figure 8: RMSE values obtained for the baseline (HSMM) and proposed (HMM-MLP) duration models for the phones in the test set.

15 sentences were randomly chosen from the test set. Each sentence was synthesised using the system described in Section 4.2 and durations obtained from the three methods respectively: baseline HSMM, proposed HMM-MLP and natural durations (measured on recorded speech).

11 subjects participated in the evaluation, 6 of whom were native speakers of English. They were asked to select the sample (A or B) of each pair (included speech synthesised using HSMM and HMM-MLP respectively) that sounded more closely to the reference speech in terms of naturalness. They were also asked to choose the third option ’X’ when they did not perceive any dif-

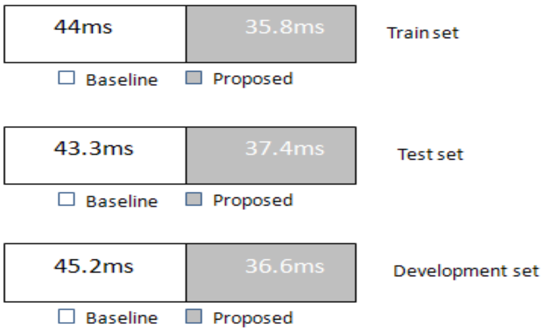

Figure 9: RMSE of phone duration averaged over all phones for the baseline and proposed duration models in the train, test and development sets respectively.

ference between the two samples.

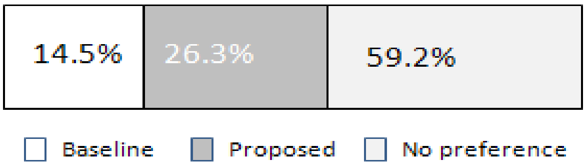

Figure 10: Preference rates for the baseline and proposed approaches.

4.5.1. *Results*

Figure 10 shows the preference rates obtained for the systems using the baseline and proposed duration modelling methods respectively. The preference rate shows that the proposed system synthesised speech that more closely resembled the reference utterance than the baseline system. Furthermore, a Friedman test was performed on the results of the perceptual evaluation to determine the statistical significance and the mean ranks for the baseline, proposed and "no preference" were 1.1, 2.4 and 2.5 respectively with a $p - value < 0.05$.

5. Conclusions

This paper presented a hybrid HMM-MLP duration modelling technique for HMM-based speech synthesis. In this approach, HMM firstly is used to obtain initial phone durations by state-level phone alignment. In the second stage MLP is used to explicitly model state duration.

An objective experiment to evaluate the hybrid HMM-MLP method for duration modelling in HMM-based speech synthesis showed that this method generally modelled more accurately phone durations as compared with a baseline system using HSMM. Furthermore, a perceptual evaluation showed that the proposed durational modelling method synthesised speech that more closely resembled the natural speech than the baseline method.

Future work will study in more details the duration modelling using HMM-MLP for the phones that obtained poor results in the objective evaluation. Also the HSMM and HMM-MLP methods will be compared in terms of speech rate transformation and for other languages and voices.

6. Acknowledgments

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/CE/I1142 as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin. 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.

Thanks to the reviewers for their useful comments on the initial version of this paper.

7. References

[1] Yamagishi, J. and Kobayashi, T. "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training", *IEICE Transaction on Information and System*, vol. E90-D, no.2, pp. 533-543, 2007.

[2] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K. and Isogai, J., "Analysis of speaker adaptation algorihms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm", *IEEE Audio, Speech, and Language Processing* vol. 17 issue 1, pp. 66-83, 2009.

[3] Hanna S., Elina, H., Jani, N. and Moncef, G., "Analysis of duration prediction accuracy in HMM-based speech synthesis", *In Proc. of the Fifth International Conference on Speech prosody*, 2010

[4] Vaseghi, S. V., "State duration modelling in hidden Markov models", *Signal Processing*, 41(1), pp. 31-41, 1995.

[5] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T., "Hidden semi-Markov model based speech synthesis", *In Proc. of INTERSPEECH*, pp. 1393-1396, 2004.

[6] Ogbureke, U. K., Cabral, J. and Carson-Berndsen, J., "Explicit duration modelling in HMM-based speech synthesis using continuous hidden Markov model", *In Proc. of the 11th International Conference on Information Sciences, Signal Processing and their Applications*, pp. 1113-1118, 2012.

[7] Shinoda, K. and Watanabe, T., "Acoustic modeling based on the MDL criterion for speech recognition", *In Proc. of Eurospeech*, pp. 99-102, 1997.

[8] Zen, H., "An example of context-dependent label format for HMM-based speech synthesis in English", *The HTS CMU-ARCTIC demo*, 2006.

[9] Sreenivasa, R. K. and Yegnanarayana, B. "Modeling durations of syllables using neural networks", *In Computer Speech and Language*, 21-2, pp. 282-295, 2007.

[10] Cordoba, R., Vallejo, J. A., Montero, J. M., Gutierrezarriola, J., Lopez, M. A. and Pardo, J. M., "Automatic modeling of duration in a Spanish text-to-speech system using neural networks", *In Proc. of the European Conference on Speech Communication and Technology*, 1999.

[11] Ben, K. and Patrick, V. D., "An introduction to neural networks", *University of Amsterdam*, 1996.

[12] Aioanei, D., "A knowledge-based and data-driven speech recognition framework", *Ph.D. Thesis, University College Dublin*, 2008.

[13] Kominek, J. and Black, A., "The CMU Arctic speech databases", *In Proc. of 5th ISCA Speech Synthesis Workshop*, pp. 223-224, 2004.

[14] Talkin, D., "A robust algorithm for pitch tracking (RAPT)", *In Speech coding and synthesis*, W. B. Klein and K. K Paliwal, (Eds.), Elsevier, pp. 495-518, 1995.

[15] Kawahara, H., Masuda-Katsuse, I. and Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds", *Speech Communication*, 27, pp. 187-207, 1999.

[16] "HMM-based speech synthesis system version 2.1", http://hts.sp.nitech.ac.jp, 2008.

[17] Nabney, I., "Netlab neural toobox", *www1.aston.ac.uk/ncrg/*, 1999.