Training a Parametric-Based LogF0 Model with the Minimum Generation Error Criterion

Javier Latorre, M.J.F. Gales, Heiga Zen

Toshiba Research Europe Ltd. Cambridge Research Laboratory, Cambridge, UK
{javier.latorre,mark.gales,heiga.zen}@crl.toshiba.co.uk

Abstract

This paper describes an approach for improving a statistical parametric-based logF0 model using minimum-generation-error (MGE) training. Compared with the previous scheme based on decision tree clustering, MGE allows the minimisation of the error in the generated logF0 to take into account not only each cluster by itself, but also the way in which the clusters interact with each other in the generation of the F0 over the whole sentence. Moreover, the "weights" of each component of the model, which previously were adjusted manually, are optimized automatically by the MGE training during the re-estimation of the model covariances. Objective evaluation indicated that, although the logF0 contours generated by the models trained with MGE have approximately the same root mean square error and correlation factor as those generated with the baseline models, they present a higher dynamic range. The subjective evaluation shows a small but significant preference for the system trained with MGE.

Index Terms: speech synthesis, prosody, statistical model, minimum generation error, fundamental frequency, F0

1. Introduction

A standard text-to-speech system consists of three modules: a front end that transforms the input text into a sequence of descriptions that can be understood by the computer; a prosody module that predicts the duration and intonation for each one of these segments; and a waveform generation module that produces the final speech signal for each segment based on their description and their predicted prosodic values. Although some unit selection systems do not need a prosody module, in statistical parametric synthesis the predicted fundamental frequency values (F0) are essential because they are needed to construct the excitation signal for the vocoder.

The goal of an intonation module is to create a trajectory of F0 values that conveys the prosodic information required by the input text as unambiguously and naturally as possible. The parametric-based F0 model [1] is a statistical method similar to standard multi-space probability distribution Hidden Markov model (HMM-MSD) F0 generation [2]. The main difference is that instead of modeling the F0 trajectory directly at a frame by frame level, F0 "contours" are modeled at well defined linguistic units such as the syllable.

Although past experiments showed that the intonation generated with this approach sounds more natural and stable than the one generated by standard HMMs [1], the baseline implementation presents several problems. First, the training consists exclusively of the clustering of the parameterized logF0 contour associated with each syllable. Such clustering minimizes the error locally for each cluster of syllables, but not globally when the syllable models are combined to generate the logF0 for the whole utterance. Second, in order to optimize the continuity of the generated logF0 values, the weights associated with the continuity component of the model have to be adjusted manually or by means of a grid search.

This paper presents a proposal to overcome these two problems by re-training the baseline model produced by a decision tree clustering with a minimum generation error (MGE) criterion [3]. The rest of the paper is organized as follows. Section 2 describes the parametric-based F0 model and its differences with other trajectory F0 models. Section 3 explains the new training algorithm based on the MGE criterion. Section 4 describes the models trained for the experiment and shows the results of an objective and subjective evaluation of the new training algorithm. Finally conclusions are drawn in section 5.

2. Parametric-based F0 model

In statistical intonation methods, the F0 or logF0 signal is considered as a random variable. For a sentence with T frames, the prediction consists of finding the vector of logF0 values $\boldsymbol{x} = [x_1, x_2, \cdots, x_T]^\top$ that maximizes its output probability

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, P(\boldsymbol{x} \mid \boldsymbol{\lambda}) \tag{1}$$

where $x_t = logF0_t$ and $\boldsymbol{\lambda}$ is a statistical model of logF0, which is usually defined as a set of Gaussians distributions.

In standard HMM-based synthesis, $\boldsymbol{\lambda}$ directly models the probability of the observed logF0 at the frame level. The main problem with this approach is that intonation acts at supra-segmental level. Statistically, this means that the logF0 values of frames belonging to the same supra-segmental unit tend to be more strongly correlated than those belonging to different units. The Δx_t and $\Delta^2 x_t$ coefficients used in standard HMMs simulates a block correlation matrix of the values within the windows used to calculate them. However, the length of such windows is fixed and its boundaries do not match any linguistic level such as phone, syllable, etc. As a result, supra-segmental information in standard HMMs appears only implicitly in the decision tree-based clustering of the state-level logF0 models.

The main idea of the parametric-based F0 model is to make supra-segmental information explicit by defining statistical models that represent the logF0 contour of supra-segmental linguistic levels. Mathematically this is equivalent to using a block-diagonal covariance matrix with variable width blocks so that each block represents the frames associated with each particular supra-segmental unit.

Define $\boldsymbol{x}^{(s)} = [x_1^{(s)}, \cdots, x_{d_s}^{(s)}]^\top$ as the logF0 vector associated with a syllable s, where d_s is the duration of the syllable in frames. If the duration of all the syllables in the training database were the same, models could be trained directly on the

$\boldsymbol{x}^{(s)}$ vectors. Unfortunately, this is not the case. Alternatively, a similar result can be obtained if a linear transformation is applied to $\boldsymbol{x}^{(s)}$ so that

$$\boldsymbol{x}^{(s)} = \boldsymbol{N}^{d_s}\boldsymbol{c}_s + \boldsymbol{\epsilon}_s \tag{2}$$

where $\boldsymbol{c}_s$ is the coefficients vector for syllable s, $\boldsymbol{N}^{d_s}$ a deterministic linear transformation which depends only on d_s, and $\boldsymbol{\epsilon}_s$ the parameterization error. In our implementation $\boldsymbol{N}^{d_s}$ is the inverse of the 5^{th} order discrete cosine transform (DCT) but other linear transformations could be used. Now, instead of modelling $\boldsymbol{x}^{(s)}$, $\boldsymbol{c}_s$ is modeled. For generation, the trajectory DCT vector $\boldsymbol{c} = [\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_S^\top]^\top$ that maximizes

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P(\boldsymbol{c} \mid \boldsymbol{\lambda}) \tag{3}$$

is first calculated, where S is the number of syllables in the sentence. If $\boldsymbol{\lambda}$ describes a Gaussian distribution and assuming that the terms of $\boldsymbol{\epsilon}_s$ in Eq. (2) also follow a Gaussian distribution with 0 mean and σ_s variance, the solution to Eq. (1) is

$$\hat{\boldsymbol{x}} = \boldsymbol{N}\hat{\boldsymbol{c}} \tag{4}$$

where $\boldsymbol{N}$ is a block diagonal matrix formed by the concatenation of the $\boldsymbol{N}^{d_s}$ matrices for each one of the S syllables in the sentence.

An additional advantage of this framework is that it allows parameterized models to be defined at different levels and combined, as exemplified in [1] and [4] with a syllable and a state level.

2.1. Concatenation coefficients

If only the DCT coefficients are modeled, $\hat{\boldsymbol{c}}$ would simply be the concatenation of the mean vectors of the Gaussian distributions associated with each syllable. In this case, the $\hat{\boldsymbol{x}}$ generated from such a DCT trajectory will not necessarily be continuous and smooth in the transition between syllables. To avoid gaps in the F0 contour, the manner in which the logF0 of one syllable relates with that of its neighbours must be modelled. This can be achieved by means of some concatenation coefficients similar to the Δ and Δ^2 features used in standard HMM-based speech synthesis [5]. To obtain a smooth spectrum the spectral envelopes of consecutive frames have to be similar to each other. However, to obtain a natural logF0 trajectory the logF0 contour of consecutive syllables does not need to be similar but the logF0 in the transition between two syllables must be continuous. The following continuity constraints are therefore applied to the model:

1. The delta of the 0^{th} DCT coefficient (Δc^0) is calculated as:
$$\Delta c_s^0 = c_{s+1}^0 - c_{s-1}^0 \tag{5}$$
which guarantees that the average logF0 of two consecutive syllables does not differ too much.
2. The gradient of logF0 at the junction with the previous and next syllables: $\Delta x_{\mathsf{b}}^{(s)}$ and $\Delta x_{\mathsf{e}}^{(s)}$. Defining the gradients in term of the logF0 values
$$\Delta x_{\mathsf{e}}^{(s)} = \sum_{w=-W}^{-1} w\boldsymbol{x}_{d_s+w+1}^{(s)} + \sum_{w=1}^{W} w\boldsymbol{x}_w^{(s+1)} \tag{6}$$
where W is a fixed window length in frames around the syllable boundary. Using the linear relationship of Eq. (2) and ignoring the $\boldsymbol{\epsilon}_s$ terms, Eq. (6) can be rewritten as
$$\Delta x_{\mathsf{e}}^{(s)} = \boldsymbol{H}_{\mathsf{e}}^{d_s}\boldsymbol{c}_s + \boldsymbol{H}_{\mathsf{b}}^{d_{(s+1)}}\boldsymbol{c}_{s+1} \tag{7}$$
where
$$\boldsymbol{H}_{\mathsf{e}}^{d_s} = \sum_{w=-W}^{-1} w\boldsymbol{N}_{w+1+d_s}^{d_s} \tag{8}$$
$$\boldsymbol{H}_{\mathsf{b}}^{d_s} = \sum_{w=1}^{W} w\boldsymbol{N}_w^{d_s} \tag{9}$$
and $\boldsymbol{N}_w^{d_s}$ is the w^{th} row of $\boldsymbol{N}^{d_s}$. Similarly
$$\Delta x_{\mathsf{b}}^{(s)} = \boldsymbol{H}_{\mathsf{e}}^{d_{(s-1)}}\boldsymbol{c}_{s-1} + \boldsymbol{H}_{\mathsf{b}}^{d_s}\boldsymbol{c}_s \tag{10}$$

2.2. Definition of the statistical model

With the addition of the continuity coefficients, the following 8-dimensional DCT observation vector $\boldsymbol{o}_s$ is obtained on which to train the model parameters $\boldsymbol{\lambda}$

$$\boldsymbol{o}_s = [\boldsymbol{c}_s^\top, \Delta c_s^0, \Delta x_{\mathsf{b}}^{(s)}, \Delta x_{\mathsf{e}}^{(s)}]^\top \tag{11}$$

As in standard HMM synthesis, $\boldsymbol{\lambda}$ consists of a set of context-dependent Gaussian distributions $\mathcal{N}(\boldsymbol{o}_s; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$. Given such a model, the probability of the DCT observation vector for a whole sentence $\boldsymbol{o} = [\boldsymbol{o}_1^\top, \cdots, \boldsymbol{o}_S^\top]^\top$ is

$$P(\boldsymbol{o} \mid \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{12}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the total mean vector and diagonal covariance matrix created by the concatenation of the $\boldsymbol{\mu}_s$ and $\boldsymbol{\Sigma}_s$ of the Gaussian distributions associated with each syllable s of the input sentence. The mapping between the syllable of an input text and their associated distribution is implemented by means of decision tree clustering.

2.3. Generation of the LogF0 contour

Given the model $\boldsymbol{\lambda}$ of the DCT observation vectors $\boldsymbol{o}_s$, the generation algorithm consists of obtaining the trajectory DCT $\hat{\boldsymbol{c}}$ that maximizes $P(\boldsymbol{c} \mid \boldsymbol{\lambda})$. All the components of $\boldsymbol{o}$ can be expressed as linear transformations of $\boldsymbol{c}_s$ for current and surrounding syllables. Therefore, as in standard HMM synthesis, $\boldsymbol{o}$ can be expressed as $\boldsymbol{o} = \boldsymbol{M}\boldsymbol{c}$, where $\boldsymbol{M}$ is a transformation matrix that depends only on the syllable durations. Since $\boldsymbol{\lambda}$ is Gaussian,

$$P(\boldsymbol{c} \mid \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{c}; \hat{\boldsymbol{c}}, \boldsymbol{P}) \tag{13}$$

where

$$\hat{\boldsymbol{c}} = \boldsymbol{L}\boldsymbol{\mu} \tag{14}$$
$$\boldsymbol{P} = (\boldsymbol{M}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{M})^{-1} \tag{15}$$
$$\boldsymbol{L} = \boldsymbol{P}\boldsymbol{M}^\top\boldsymbol{\Sigma}^{-1} \tag{16}$$

The value that maximizes $P(\boldsymbol{c} \mid \boldsymbol{\lambda})$ is obviously $\hat{\boldsymbol{c}}$. Finally, the logF0 vector $\hat{\boldsymbol{x}}$ is calculated from Eq. (4).

This F0 generation process resembles the one proposed in [6] in the sense that it maximizes a function formed by the summation of a "disambiguation cost", defined by the DCT coefficients of each syllable logF0 contour, and an "ease of production cost", defined by the continuity coefficients. The main difference is that in this model, both costs are defined as likelihoods. Therefore their weights can be obtained directly and automatically from the training material.

3. Training with minimum generation error

In the previous implementation [1], the training process consisted of clustering the syllable observation vectors $\boldsymbol{o}_s$ using a decision tree. This clustering maximizes the local likelihood at each split. However, it does not necessarily minimize the error over the generated $\hat{\boldsymbol{c}}$, and associated $\hat{\boldsymbol{x}}$, once the concatenation coefficients are considered. In order to improve the model, the approach proposed in [3] is adopted of minimizing the total error between the generated logF0 and the observed logF0 over the whole training data. Let u specify the utterance index, so that, for example, the observed and predicted logF0 vectors over the complete utterance are called $\boldsymbol{x}^{(u)}$ and $\hat{\boldsymbol{x}}^{(u)}$ respectively. Define the error for one utterance as the weighted Euclidean distance between $\boldsymbol{x}^{(u)}$ and $\hat{\boldsymbol{x}}^{(u)}$. The total generation error on the training data as a function of $\boldsymbol{\lambda}$ is

$$\mathcal{F}(\boldsymbol{\lambda}) = \sum_{\forall u} (\boldsymbol{x}^{(u)} - \hat{\boldsymbol{x}}^{(u)})^\top \boldsymbol{W}_u (\boldsymbol{x}^{(u)} - \hat{\boldsymbol{x}}^{(u)}) \tag{17}$$

where $\boldsymbol{W}_u$ is an appropriate error weighting matrix. A selection matrix $\boldsymbol{A}_u$ is defined according to the decision tree so that for the utterance u the mean and variance are obtained as

$$\boldsymbol{\mu}_u = \boldsymbol{A}_u \boldsymbol{\eta} \tag{18}$$

$$\boldsymbol{\Sigma}_u = \text{diag}(\boldsymbol{A}_u \boldsymbol{\upsilon}) \tag{19}$$

where $\boldsymbol{\eta}$ is the supervector made of the concatenation of all the mean vectors of $\boldsymbol{\lambda}$, $\boldsymbol{\upsilon}$ the supervector made of the concatenation of the leading diagonal of all the covariance matrices of $\boldsymbol{\lambda}$, and diag a function that creates a square matrix with the input vector as leading diagonal. The generated logF0 vector can be rewritten using Eq. (4) as

$$\hat{\boldsymbol{x}}^{(u)} = \boldsymbol{N}_u \hat{\boldsymbol{c}}_u = \boldsymbol{N}_u \boldsymbol{L}_u \boldsymbol{A}_u \boldsymbol{\eta} \tag{20}$$

Now, the derivative of Eq. (17) with respect to $\boldsymbol{\eta}$ is

$$\frac{\partial \mathcal{F}(\boldsymbol{\lambda})}{\partial \boldsymbol{\eta}} = 2(\boldsymbol{Q}\boldsymbol{\eta} - \boldsymbol{q}) \tag{21}$$

where

$$\boldsymbol{Q} = \sum_{\forall u} \boldsymbol{A}_u^\top \boldsymbol{L}_u^\top \boldsymbol{N}_u^\top \boldsymbol{W}_u \boldsymbol{N}_u \boldsymbol{L}_u \boldsymbol{A}_u \tag{22}$$

$$\boldsymbol{q} = \sum_{\forall u} \boldsymbol{A}_u^\top \boldsymbol{L}_u^\top \boldsymbol{N}_u^\top \boldsymbol{W}_u \boldsymbol{x}^{(u)} \tag{23}$$

Therefore, the mean value that minimizes the error is

$$\hat{\boldsymbol{\eta}} = \boldsymbol{Q}^{-1} \cdot \boldsymbol{q} \tag{24}$$

The derivative of $\mathcal{F}(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\upsilon}^{-1}$ is

$$\frac{\partial \mathcal{F}(\boldsymbol{\lambda})}{\partial \boldsymbol{\upsilon}^{-1}} = \sum_{\forall u} \left(\boldsymbol{A}_u^\top \boldsymbol{\Gamma}_u \boldsymbol{M}_u \boldsymbol{P}_u \boldsymbol{N}_u^\top \boldsymbol{W}_u \hat{\boldsymbol{e}}_u \right) \tag{25}$$

where

$$\boldsymbol{\Gamma}_u = \text{diag}(\hat{\boldsymbol{\mu}}_u - \boldsymbol{M}_u \boldsymbol{L}_u \boldsymbol{A}_u \hat{\boldsymbol{\eta}}) \tag{26}$$

$$\hat{\boldsymbol{e}}_u = \boldsymbol{x}^{(u)} - \boldsymbol{N}_u \boldsymbol{L}_u \boldsymbol{A}_u \hat{\boldsymbol{\eta}} \tag{27}$$

Equation (25) has no closed-form solution and has to be solved by iterative methods.

Once the variance is updated, it can be used to re-estimate the mean value, continuing the process iteratively.

4. Experiment

4.1. Experimental settings

The models were trained on a speech database of 4639 sentences and approximately 4.5 hours of speech from a single American English female speaker. The number of syllables in this database was 65129. The database was automatically annotated both phonetically and syntactically. In addition to the usual decision tree questions about phonetic and syntactic information, questions were also included about the duration of the syllable, its head and its coda. Previous experiments showed that these kind of questions help to model the interrelation between logF0 and duration [7]. During synthesis the duration features were obtained from the duration predicted by an HMM model.

In order to calculate the DCT of the syllable logF0 contours, the observed logF0 was interpolated with a spline function. The spline interpolator was set to minimize an error calculated with respect to 5 points at each edge of the discontinuity. The error was weighted so that the error tolerance at the immediate edges of the discontinuity was 0.05 semitones and doubled for each frame moving away from the edge. To reduce the effect of F0 extraction errors on the interpolated logF0 values, the interpolation was calculated using only "reliable" values, which were selected as those fulfilling all the following heuristic criteria:

a) values classified as voiced by the F0 extraction algorithm;

b) values that present an autocorrelation value higher than a threshold, which was set manually to 85%;

c) values that belong to phones that present a clear periodicity, i.e., vowels, semivowels and nasals; and

d) for syllables with a total dynamic range higher than 8 semitones, values within 1.65σ of the central value.

Finally, any accepted value that after the selection becomes surrounded by non-reliable values was also considered non-reliable and ignored for the interpolation.

The interpolated logF0 segments, $\boldsymbol{x}^{(s)}$, were chunked according to the syllable boundaries given by the phone alignment. Since the interpolation is calculated only using acceptable values and all utterances start and finish with silence, the beginning of the first syllable and the end of the last one might not have a valid interpolated F0 value. To reduce this problem, the acceptance criteria for the first and last syllables were relaxed so that they were accepted if they fulfil criterion (a) and belong to a voiced phone. For those inital and final syllables that still started or ended with non-interpolated values, their boundaries were rearranged to match the position of the first and last accepted logF0 value in the utterance, respectively.

The syllable observation vectors were clustered with a decision tree using HTS [8] and the maximum likelihood criterion. For this purpose, each syllable was modelled as a single-state model, with its mean set to the observation $\boldsymbol{o}_s$ and its variance set to $\boldsymbol{0}$. The variance floor was set to the variance of $\boldsymbol{o}_s$ over the training data scaled by 10^{-5}.

For syllables immediately before or after silence the values of the continuation coefficients $\Delta x_{\text{b}}^{(s)}$ and $\Delta x_{\text{e}}^{(s)}$ have no real meaning. To avoid mixing them with other "real" values in the clustering, these two coefficients were defined as MSD streams. In this way, if the border of the syllable is a silence, the weight of the real mean value Gaussian of the pseudo-model is set to 0 and the weight of the 0-dimensional value is set to 1. For the F0 generation, the variances of$\Delta x_{\text{b}}^{(s)}$ and $\Delta x_{\text{e}}^{(s)}$ are set to infinite if the syllable s is preceded or followed by a silence,

respectively, or if the weight of the 0-dimensional Gaussian is above a threshold, which was set manually to 0.5

The state occupancy of all syllables were given an occupancy value of 1. This shape-based approach was found to produce better results than the error-based one used in [1] where each syllable was assigned an occupancy according to its duration. Finally, for the experiment the MDL threshold was set to 0.4 which produced a total of 1076 final leaves. For the MGE training the means and variances were updated once.

Due to the MSD definition of the $\Delta x_{\mathsf{b}}^{(s)}$ and $\Delta x_{\mathsf{e}}^{(s)}$ coefficients, some of the rows of the Hessian matrix $\boldsymbol{Q}$ in Eq. (21) are all zeros. Therefore, to calculate $\hat{\boldsymbol{\mu}}$ directly, it is convenient to transform Eq. (24) so that all those obvious null dimensions of $\boldsymbol{Q}$ are removed. Additionally, due to the linear relationship between the DCT and the continuity parameters described in section 2.1, the mean value of the concatenation parameters for those clusters that contain only one syllable can become a linear combination of the DCT coefficients of those syllables. In such cases $\boldsymbol{Q}$ again becomes singular. To avoid this problem, the clustering was configured so that each leaf contains at least 3 examples. In any case, it can be difficult to obtain a direct solution to Eq. (24) due to the high dimensionality of $\boldsymbol{Q}$. In such cases, an iterative algorithm such as conjugate gradient can be more efficient.

The goal of the error weighting matrix is to further reduce the error on the reliable observed logF0 data by allowing a greater tolerance for the interpolated ones. Different approaches were tried to define this weighting matrix. However, to date none of these approaches have improved performance over a baseline identity matrix.

4.2. Objective evaluation

To obtain an initial idea of the effectiveness of the MGE training, some objective measurements were calculated on a held-out test subset of 511 sentences. Table 1 shows the results. On the training data, MGE training reduced the RMSE and improved the correlation. On the testing data, the RMSE of MGE models is actually worse. However, the correlation is basically the same, and the average dynamic range was larger.

Table 1: *Objective evaluation of the generated logF0.*

	RMSE (semitones)		Correlation coefficient		Dynamic range (semitones)	
	train	test	train	test	train	test
Data	-	-	-	-	16.21	16.05
Baseline	2.88	2.85	0.75	0.71	12.39	12.33
MGE	2.58	3.08	0.77	0.71	13.39	13.54

4.3. Subjective evaluation

In order to see whether the larger dynamic range correlates with perceived quality, a subjective preference test was run. For synthesis, the spectrum, duration and aperiodicity generated by an standard HMM model trained on the same data was used. The DCT F0 model always generates a continuous logF0 trajectory. Since the standard HMM source excitation requires the F0 of unvoiced frames to be 0, the generated F0 was "de-voiced" using the voice/unvoiced information of the F0 trajectory generated by the standard HMM-MSD model. This voicing signal was also used to adjust the duration of the first and last syllable of each sentence.

In the evaluation, 11 subjects were presented with 45 stimuli pairs randomly selected from a pool of 220 sentences. The order of each pair was also random. Subjects were asked to choose the stimulus they judged better, or "None" if they could not find any difference. All subjects were speech technology experts. Six of the subjects were native English speakers, and the other 4 highly proficient English speakers who have been living in English speaking countries for several years. The length of each evaluation sentence was between 7 and 14 words, with an average of 9.5 and a mode of 9 words.

Table 2 show the general results of the evaluation. Although for most sentences no difference was perceived, the preference for MGE and baseline models were still significantly different. Overall, all but one subject preferred the stimuli with the F0 generated by the MGE trained model. On a sentence basis, the MGE-trained model was preferred for 35.9% of sentences versus 17.7% for the baseline. 46.4% of sentences had exactly the same number of votes for both models. No obvious preferences were observed for one system over the other for specific types of sentences.

Table 2: *Results for all subjects.*

Baseline training	None	MGE training	Difference 95% margin	Binary z-score error
14.7%	57.6%	27.7%	$13\% \pm 5.85\%$	0.19%

5. Conclusion

In this paper a new training method has been presented for a parametric-based F0 model based on the minimum generation error criterion. Whereas the previous training optimized the model locally for each cluster, the new MGE training optimizes all the clusters together taking into consideration the way in which they interact with each other to generate the logF0 for the whole utterance. Objective tests showed that MGE produced an improvement in the dynamic range of the synthesized F0 contour, which resulted in a small but significant subjective preference. The new training scheme also eliminates the need to adjust manually the weights of individual model components.

6. References

[1] J. Latorre, M. Akamine, "Multilevel parametric-base F0 model for speech synthesis", in Proc. Interspeech, Brisbane, 2008.

[2] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling" in Proc. ICASSP, Phoenix, 1999.

[3] Y.J. Wu, R.H. Wang, "Minimum generation error training for HMM-based speech synthesis" in Proc. ICASSP, Toulouse, 2006.

[4] Y. Qian, Z. Wu, F.K.Soong "Improved prosody generation by maximizing joint likelihood of state and longer units" in Proc. ICASSP, Taipei, 2009.

[5] K. Tokuda, T. Kobayashi, S. Imai, "Speech parameter generation from HMM using dynamic features". in Proc. ICASSP, Detroit, 1995.

[6] G.P. Kochanski and C. Shih, "Stem-ML:Language-Independent prosody description" in Proc. ICSLP, Beijing, 2000.

[7] J. Latorre, S. Buchholz, M. Akamine, "Usage of an external duration model for HMM-based speech synthesis" in Proc. 5^{th} Speech Prosody Workshop, Chicago, 2010.

[8] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, "The HMM-based Speech Synthesis System (HTS) version 2.0". in Proc. SSW6, Bonn, 2006.