INTERSPEECH 2009 BRIGHTON

Feedback Loop for Prosody Prediction in Concatenative Speech Synthesis.

Javier Latorre[1], *Sergio Gracia*[2], *Masami Akamine*[1]

[1]Toshiba Corporate Research & Development Center, Japan
[2]TelecomBCN, Universitat Politecnica Catalunya, Spain
{javier.latorre,masa.akamine}@toshiba.co.jp, sgra8435@alu-etsetb.upc.edu

10.21437/Interspeech.2009-593

Abstract

We propose a method for concatenative speech synthesis that permits to obtain a better matching between the logF0 and duration predicted by the prosody module and the waveform generation back-end. The proposed method is based upon our previous multilevel parametric F0 model and Toshiba's plural unit selection and fusion synthesizer. The method adds a feedback loop from the back-end into the prosody module so that the prosodical information of the selected units is used to re-estimate new prosody values. The feedback loop defines a frame-level prosody model which consists of the average value and variance of the duration and logF0 of the selected units. The log-likelihood defined by this model is added to the log-likelihood of the prosody model. From the maximization of this total log-likelihood, we obtain the prosody values that produce the optimum compromise between the distortion introduced by F0 discontinuities and the one created by the prosody adjusting signal processing.

Index Terms: speech synthesis, multilevel, parametric F0, prosody, Discrete cosine transform, log-likelihood maximization

1. Introduction

A good prosody model is one of the factors that contributes more decisively to the overall quality and naturalness of a speech synthesizer. It is often observed that while a good prosody can not make a bad synthetic voice sound good, a bad one can make an otherwise good synthesizer sound badly. In unit selection systems, the values estimated by the prosody model are used in two ways: first, to select the units from the database, and second, to modify the pitch and duration of the selected units. These modifications however, introduce a distortion proportional to the degree of signal processing which is required. One possible solution to eliminate such distortion is not to apply any modification at all and rely on the natural pitch and duration of the selected units. This approach works reasonably well in systems built on huge amount of data ($\geq$ 8 hours). In this case, units are often contiguous in the database and little or no modification is actually needed. In smaller databases however, the selected units are rarely contiguous. If we concatenate them 'as-they-are', we are likely to create chunks and discontinuities in the synthesized speech which are more annoying than the signal processing distortion we are trying to avoid.

Several methods have been proposed to combine the intonation values predicted by the prosody model with those of the database. In [1, 2], after the units are selected, a set of rules are used to decide the degree to which their duration and pitch should be modified. Though this method helps to reduce the distortion while keeping a sufficient coherence for the prosody, the definition of the rules is not simple. For a limited domain synthesizer, a more dynamic approach was proposed in [3]. There, a weighted finite-state transducer framework is used to jointly predict the prosody and select the units. A similar method was proposed more recently, by Campillo et al. [4]. Their method simultaneously selects the synthesis units and the pitch contour models out of a set of pitch contour candidates. The main drawback of this method is its high computational cost.

The method that we propose is similar to the methods mentioned above in the sense that we attempt to optimize simultaneously the unit selection and the pitch contour generation. The main difference is that instead of a set of prosodic templates (pre-computed [3] or created on-line from the database in a previous step [4]), we use a continuous pitch contour model based on the parameter generation algorithm [5]. This allows us to obtain good generalization with a limited computational cost and a minimal footprint. In our implementation we use a parametric multilevel prosody model [6] and a plural unit selection and fusion synthesizer [9]. A brief introduction to these two techniques is provided in sections 2 and 3 respectively. The details of our approach are explained in section 4. Section 5 introduces the results of some preliminary evaluation and finally in section 6 we draw the conclusions.

2. Parametric multilevel F0 model

The parametric multilevel F0 model is a method to generate a pitch contour for a sentence using statistical models of the pitch contours at one or more linguistic levels, such as the phone, syllable, or phrase. In order to train these models, the pitch contours of the units in the database are first parameterized, so that for each level l, they can be represented by vectors $\boldsymbol{o}^l$, all with the same dimension. These vectors are then grouped into clusters, for example by means of a decision tree, from which sufficient statistics are obtained. In our current implementation the sufficient statistics are calculated directly and independently for each cluster and each level. The synchronization between levels is achieved by using the same phone-level segmentation for all levels. A joint re-training of all the levels with, for example, a minimum generation error algorithm [7] is also possible.

At the synthesis stage, the input sentence is converted into a sequence of statistical models from where we define a log-likelihood function $F^l(\boldsymbol{o}^l)$ for each linguistic level l. Next, these log-likelihood functions are expressed in terms of the same set of parameters that describe the pitch contour at the main level $\boldsymbol{x}^m$ as

$$F^l(\boldsymbol{o}^l) = F^l\big(G^l(\boldsymbol{x}^m)\big) \tag{1}$$

and integrated as a weighted summation into a global log-

likelihood function:

$$F(\boldsymbol{x}^m) = \sum_{\forall l} \lambda_l F^l\big(G^l(\boldsymbol{x}^m)\big) \quad (2)$$

with λ_l the weight of the l-level. This function is maximized with respect to $\boldsymbol{x}^m$ and finally, the logF0 pattern is obtained from the inverse transformation of $\boldsymbol{x}^m$ according to the estimated duration of the units at the main level.

The main advantage of this probabilistic framework is that models at different linguistic levels can be easily integrated to explicitly model prosodic effects such as emphasis, questioning intonation, etc, at the level where the impact is stronger. The only requirement to integrate a new level is for its $G^l(\boldsymbol{x})$ to have an analytic closed formulation.

2.1. Parameterization

In our implementation the main linguistic level is the syllable and the parameters $\boldsymbol{x}^m$ are the first 5^{th} coefficients of the discrete cosine Transform (DCT) of the syllable pitch contour. The process to calculate these coefficients is as follows: first, the pitch contour of the utterance is interpolated; then, the sections of pitch associated to each syllable are chunked; and finally, the parameterization is applied to each chunked segments in the utterance.
In the synthesis, the continuity of the generated pitch is achieved by two means: the continuity parameters included in some levels, and the interaction between levels. The continuity parameters included at some levels describe the relationship between the pitch contour of one unit and those of its neighbors. In this way, they impose a constraint on how much the pitch contour can vary from one unit to the next one. The most important continuity constraint in our implementation is $\Delta LogF0$ at syllable level, which represents the gradient of the LogF0 at the junction points between two syllables. The addition of the log-likelihood of any new level acts also as a set of constraints over the values of $\boldsymbol{x}^m$. The main additional level in our implementation is an accent group level model as described in [8], which imposes continuity restrictions on the values of the 0^{th} coefficient of $\boldsymbol{x}^m$.

3. Plural unit selection and fusion

The plural unit selection and fusion method [9] is a speech synthesis method that combines the naturalness of standard unit selection methods with the robustness of unit-training methods [10]. In this method, instead of a single unit, a cluster of U speech units are selected from the database for each unit of the input text. Typical values for U are between 3 and 10 speech units. Figure 1 illustrates the way in which the unit clusters are selected. First, we find the path that minimize a total cost function made of a target sub-cost and a concatenation sub-cost. This step is identical as in standard unit selection synthesis systems. Then, for each target unit, another $U-1$ samples are selected. These samples are those with lower target cost and lower concatenation cost with respect to the neighbor units of the optimum path. Next, the U samples of each cluster are fused together in time domain to create a single fused unit. The pitch and duration of the fused units are then modified according to the input prosody values and overlap-added to generate the speech waveform.

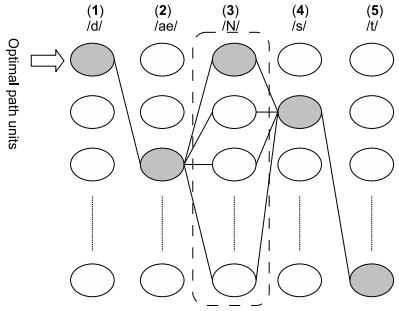

Figure 1: *Plural unit selection process [9].*

4. Addition of a feedback loop

In comparison with standard unit selection, the plural unit selection and fusion method reduces largely the spectral distortion produced by the F0 modifications. However, this distortion does not disapear completely. A possible way to further reduce it consists in using the prosody of the fused units or that of the units in the optimum path. However, neither the pitch of the optimum path, nor the average pitch contours of the fused units are continuous in most cases. Therefore additional smoothing over the pitch contour of the fused-units has to be applied to eliminate the gaps.

In our proposal, instead of reducing the gaps by applying a man-made smoothing to the natural pitch contour of the units, we integrate the units' natural prosody into the statistical framework described in section 2. Figure 2 depicts the block diagram of our method. First the prosody module makes an initial estimation for the duration and pitch contour of the units. Using this initial estimation, the multiple-unit-selection module chooses from the database a cluster of samples for each target unit as described in section 3. From this cluster, the statistics of the pitch contour and the duration of the units are calculated. The log-likelihood defined by these statistics, is equivalent to the log-likelihood defined by a frame-level model. Therefore it can be integrated into the total log-likelihood function of the prosody model as an additional set of constraints. The weight of the feedback model can be based directly on the variance of the pitch contour in each cluster of units. Next, a new pitch contour is obtained from the maximization of the total log-likelihood function. This new contour can be used either to modify the previously selected units, or to select new clusters of units from the database, from which to calculate a new pitch contour, repeating this process until the optimum fused units for the input text are found.

To create the feedback model, the pitch contour and duration of each unit have to be expressed in such a way that sufficient statistics can be easily computed, and that they can be expressed in terms of the primary parameters of the prosody module $\boldsymbol{x}^m$.

4.1. Feedback model for duration

In our multilevel parametric logF0 model, the estimation of the duration precedes that of the pitch contour. The main level for the duration estimation is the phone.

Synthesis units are commonly defined at sub-phonetic level

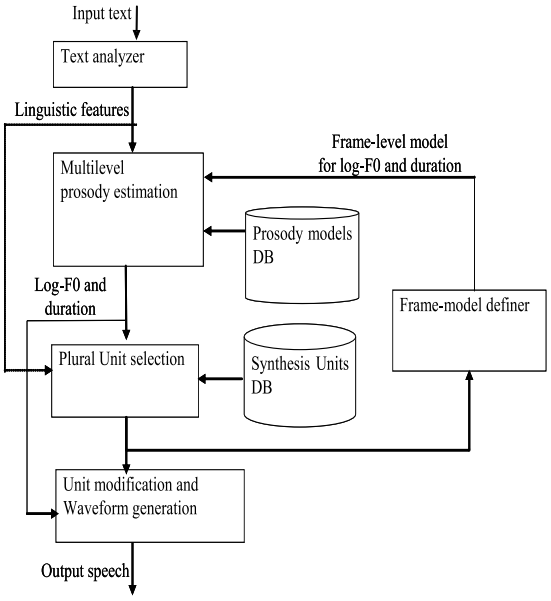

Figure 2: *Schematic diagram of the speech production.*

such as half-phones or states. But as long as phones are defined as sets of such units, the integration of the feedback duration is straight forward.

4.2. Definition of the feedback LogF0 model

The definition of a feedback model for the pitch contour is slightly more complex than for the duration because it admits more possible variants. The variant to use will depend on the amount of available memory and/or computational power for the given application. Assuming a single Gaussian distribution, we can derive the general form of the feedback log-likelihood for the pitch for all the fused half-phones hp that belong to a given syllable s as follows,

$$F^{fbck} = \sum_{\forall s} \sum_{\forall hp \in s} F_{hp}^{fbck}(\boldsymbol{o}_{hp}) \quad (3)$$

$$F_{hp}^{fbck}(\boldsymbol{o}_{hp}) = \frac{-1}{2}\big(\boldsymbol{o}_{hp} - \boldsymbol{\mu}_{hp}\big)^\top \boldsymbol{\Sigma}_{hp}^{-1}\big(\boldsymbol{o}_{hp} - \boldsymbol{\mu}_{hp}\big) + Const \quad (4)$$

where *const* is a constant and $\boldsymbol{o}_{hp}$, $\boldsymbol{\mu}_{hp}$ and $\boldsymbol{\Sigma}_{hp}$ are respectively the parameterized vector, mean value and covariance of the pitch contour of the hp half-phone. The easiest way to define $\boldsymbol{o}_{hp}$ is to use a linear transformation of the pitch contour so that:

$$\boldsymbol{o}_{hp} = \boldsymbol{H}_{hp} \cdot \boldsymbol{logF0}_{hp} = \boldsymbol{H}_{hp} \cdot \boldsymbol{S}_{hp} \cdot \boldsymbol{logF0}_{s} \quad (5)$$

where $\boldsymbol{logF0}_{hp}$ and $\boldsymbol{logF0}_{s}$ are respectively the pitch contour of the hp unit and the syllable s to which hp belongs, $\boldsymbol{H}_{hp}$ is the transformation matrix, and $\boldsymbol{S}_{hp}$ a matrix that extracts $\boldsymbol{logF0}_{hp}$ from $\boldsymbol{logF0}_{s}$.

To integrate this log-likelihood into Eq. (2), we need to express it in terms of the primary parameter vector of the syllable level $\boldsymbol{x}^m$, which is defined as

$$\boldsymbol{x}^m = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_S^\top]^\top \quad (6)$$

$$\boldsymbol{x}_s = \boldsymbol{T}_s \cdot \boldsymbol{logF0}_{s} \quad (7)$$

where $\boldsymbol{T}_s$ is the transformation matrix that corresponds to the first 5 coefficients of the DCT. Since the DCT is an invertible transformation, Eq. (5) can be expressed as

$$\boldsymbol{o}_{hp} = \boldsymbol{H}_{hp} \cdot \boldsymbol{S}_{hp} \cdot \boldsymbol{T}_s^{-1} \cdot \boldsymbol{x}_s = \boldsymbol{M}_{hp} \cdot \boldsymbol{x}_s \quad (8)$$

Therefore Eq. (4) can be rewritten as

$$F_{hp}^{fbck} = \frac{-1}{2}\big(\boldsymbol{M}_{hp}\cdot\boldsymbol{x}_s - \boldsymbol{\mu}_{hp}\big)^\top \boldsymbol{\Sigma}_{hp}^{-1}\big(\boldsymbol{M}_{hp}\cdot\boldsymbol{x}_s - \boldsymbol{\mu}_{hp}\big) + Const \quad (9)$$

and the differential of F^{fbck} with respect to $\boldsymbol{x}^m$ becomes

$$\frac{\partial F^{fbck}}{\partial \boldsymbol{x}^m} = \sum_{\forall s} \big(\boldsymbol{A}_s \cdot \boldsymbol{x}_s + \boldsymbol{B}_s\big) \quad (10)$$

with

$$\boldsymbol{A}_s = \sum_{\forall hp \in s} \boldsymbol{M}_{hp}^\top \boldsymbol{\Sigma}_{hp} \boldsymbol{M}_{hp} \quad (11)$$

$$\boldsymbol{B}_s = \sum_{\forall hp \in s} \boldsymbol{M}_{hp}^\top \boldsymbol{\Sigma}_{hp} \boldsymbol{\mu}_{hp} \quad (12)$$

From Eq. (8) we can see the matrix $\boldsymbol{H}_{hp}$ is the one that defines the feedback model. . Using $\boldsymbol{H}_{hp}$, the sufficient statistics $\boldsymbol{\mu}_{hp}$ and $\boldsymbol{\Sigma}_{hp}$ of the fused unit are calculated from the cluster of U units as follows:

$$\boldsymbol{\mu}_{hp} = \frac{1}{U}\sum_{u=1}^{U} \boldsymbol{H}_u \cdot \boldsymbol{logF0}_u \quad (13)$$

$$\boldsymbol{\Sigma}_{hp} = \frac{1}{U}\sum_{u=1}^{U} \big(\boldsymbol{H}_u \cdot \boldsymbol{logF0}_u\big)\big(\boldsymbol{H}_u \cdot \boldsymbol{logF0}_u\big)^\top - \boldsymbol{\mu}_{hp} \cdot \boldsymbol{\mu}_{hp}^\top \quad (14)$$

In general, once we have decided the transformation, $\boldsymbol{H}_u$ depends only on D_u, the length of $\boldsymbol{logF0}_u$. Two possible ways to implement this matrix are a point-based approach and a parametric approach.

In a point-based approach, the model describes the mean value and variance of the logF0 at certain points of the fused unit, for example, the beginning, middle and end of each half-phone. In this case $\boldsymbol{H}_u$ would be a $3 \times D_u$ matrix, with ones at the positions $(1, 1)$, $(2, D_u/2)$ and $(3, D_u)$ and 0 otherwise. The integration of this type of models with the syllable DCT is basically the same as the integration of a state-based HMM F0 model described in [6].

In a parametric approach, the model does not describe the exact value of logF0 at any given point, but the values of a linear parameterization of the whole pitch contour of the half-phone, for example, by means of the first 2 or 3 coefficients of the DCT of $logF0_u$. The advantage of the parametric approach is that it can provide better description of the pitch contour with less data. It also provides an easier way to mix the pitch contour of the units cluster. The disadvantages is that it requires additional memory or more computational time to calculate the parameters on-line.

5. Subjective evaluation

Although a formal evaluation has not been completed yet, an internal informal evaluation already revealed several interesting results. In this evaluation, 12 subjects compared in two pair tests utterances synthesized with and without feedback. In the first test, the stimuli consist of 25 English sentences generated from

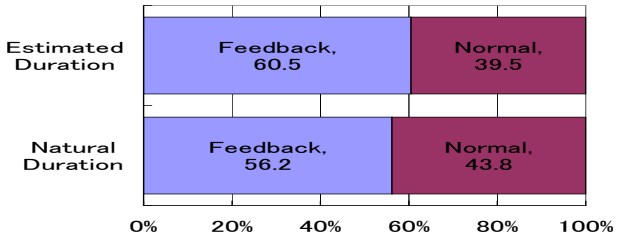

Figure 3: *Preference for the feedback model for estimated and natural phone duration.*

a text input for which the phone duration was estimated with the Quantification Method Type 1 (QMT1) [11]. In the second one, the stimuli were another 25 English sentences from a spared sub-set of our database for which the duration was copied from the natural one spoken by the voice talent. The pairs of stimuli were presented to the subjects randomly, in such a way that at the end of the evaluation each subject listened each pair of stimuli twice, each time with a different order. The database used to train the prosody model and the unit selection and fusion synthesizer was the same as the one described in [12]. The implementation of the feedback F0 model was a point-based approach using the initial and final logF0 value of voiced half-phones. The number of clustered units, U, was 10.

5.1. Results

Figure 3 shows the average preference of these tests. In general, the addition of a feedback model contributes to improve the naturalness of the synthesized speech.

In the sentences generated with estimated duration, we obtained a statistically significant global preference for the model with feedback of 60.5% ($p < 1.0e^{-6}$), with 95% confidence interval by utterance of $\pm 4.6\%$ and by subject of $\pm 4\%$. By genre, long sentences and sentences from car-navigation domain showed the highest preference, 66% ($p < 1.0e^{-6}$) and 64% ($p < 1.0e^{-3}$) respectively. For short sentences and yes/no questions the feedback model was also preferred in 53% of the cases, but the differences were not significant.
As expected, the differences were not that clear for the test with sentences generated with natural duration. Nevertheless, there was a significant preference for the model with feedback of around 56% ($p < 5.0e^{-3}$), with 95% confidence intervals by utterance of $\pm 3.6\%$ and by subject of $\pm 6.9\%$. By genres, the higher preferences were for long sentences, 58% and questions 61%, both significant ($p < 1.0e^{-2}$). In the other two genres, short sentences and exclamations, the preference was 51% and 53% respectively, but not statistically significant.

5.2. Discussion

From the results of the preference tests and from an interview with the subjects, we think that the main improvement in our current implementation comes from the modification of the fused-units duration. At the moment, the duration after adding the feedback is directly the average duration of the fused units. In addition to reducing the signal processing distortion, the feedback duration also improves the rhythm of the utterance. That is probably why the preference for the feedback model for long sentences is higher than average with both estimated and natural duration.

The clearest effect of the feedback pitch model is the shifting of the average logF0 of the utterance to a level adequate for the sentence. In the model without feedback, the synthesized logF0 is shifted by default to the pre-computed average logF0 of the speaker. The feedback loop makes this average pitch varies from sentence to sentence. This was especially clear in the set generated with natural duration for question sentences which, at least for our voice talent, usually present higher average logF0 than other type of sentences.

6. Conclusions

We have presented a new method to get a better integration between the prosody estimation and the waveform generation back-end. This method consists in a feedback loop from the back-end to the prosody model with information about the statistics of the prosody of the selected units. The prosody model uses the feedback information to re-estimate new prosody values. The proposed method improves the quality of synthetic speech, especially for long sentences with estimated duration, though improvements can be appreciated also in the case of using natural speech duration. It can be expected that a more elaborated implementation of the algorithm will result in further reduction of the distortion and thus in synthesized speech with higher quality and naturalness.

7. Acknowledgements

The authors wish to thank the members of Toshiba Research Europe Limited, Cambridge, for their help and patience in the evaluation of the samples.

8. References

[1] Kondo et al. NEC, "Speech synthesis apparatus" Patent US 6,405,169 Jun 2002

[2] P. Taylor, "Concept-to-Speech synthesis by phonological structure matching" Philosophical Transcations of the Royal Society, 2000, Series A. 356 (1769):1403-1416

[3] Bulyko, I., and Ostendorf, M., "Joint prosody prediction and unit selection for concatenative speech synthesis" Proc. ICASSP 2001

[4] Campillo, F., and Rodriguez, E., "A method for comnining intonation modelling and speech unit selection in corpuse-based speech synthesis systems" Speech communication, v. 48, pp. 941-956, 2006

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

[6] Latorre, J. and Akamine, M., "Multilevel parametric-based F0 model for speech synthesis" Proc. Interspeech 2008

[7] Wu, Y., and Wang, R., "Minimun generation error training for HMM-based speech synthesis" Proc. ICASSP 2006

[8] Gracia, S., Latorre, J., and Akamine, M., "Inclusion of an accent group level in a multilevel parametric-base F0 model" Proc. Spring Meeting of the Acoustic Society of Japan, 2009, 1-6-2

[9] Mizutani, T. and Kagoshima, T. "Concatenative speech synthesis based on the plural unit selection and fusion method". IEICE Trans., vol. E88-D, no.11, 2005, pp.2565-2572.

[10] Akamine, M., and Kagoshima, T., "Analytic generation of synthesis units by closed loop training for tatally speaker driven text to speech system (TOS drive TTS)" Proc. ICSLP 1998

[11] Hayashi, C, "On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view". Ann. the Institute of Statistical Mathematics Vol 3, no 2, pp.69-98, 1952.

[12] Krstulovic, S., Latorre, J., Buchholz, S., "Comparing QMT1 and HMMs for the synthesis of American English prosody" Proc.Speech Prosody 2008, Campinas, Brazil