TRAINING A SUPRA-SEGMENTAL PARAMETRIC F0 MODEL WITHOUT INTERPOLATING F0

Javier Latorre[†]*,Mark J.F. Gales*[†]*,Kate Knill*[†*]*, Masami Akamine*[‡]

[†]Toshiba Research Europe Ltd., Cambridge Research Laboratory, Cambridge, UK
[‡]Toshiba Corporate Research & Development Center, Kawasaki, Japan
javier.latorre@crl.toshiba.co.uk

ABSTRACT

Combining multiple intonation models at different linguistic levels is an effective way to improve the naturalness of the predicted F0. In many of these approaches, the intonation models for supra-segmental levels are based on a parametrization of the log-F0 contours over the units of that level. However, many of these parametrisations are not stable when applied to discontinuous signals. Therefore, the F0 signal has to be interpolated. These interpolated values introduce a distortion in the coefficients that degrades the quality of the model. This paper proposes two methods that eliminate the need for such interpolation, one based on regularization and the other on factor analysis. Subjective evaluations show that, for a Discrete-cosine-transform (DCT) syllable-level model, both approaches result in a significant improvement w.r.t. a baseline using interpolated F0. The approach based on regularization yields the best results.
Index Terms: speech synthesis, intonation, factor analysis, regularization, F0 interpolation

1. INTRODUCTION

Intonation is the temporal variation of pitches. It is an essential part of speech for all human languages, which use it to encode a variety of information such as the type of sentence (question,statement), word emphasis, discourse structure, etc. Most of the information encoded in the intonation is supra-segmental. This means that its structures are at a linguistic level higher than the phone. In that sense, intonation should be considered to be continuous and smooth, at least over the time scales defined by those supra-segmental structures [1].

A problem to create an intonation model is that the pitch is a subjective psychoacoustical property of sound which cannot be obtained directly from the waveform. Instead, the fundamental frequency (F0) is used as its closest measurable proxy. However, F0 does not exist or is unobservable for unvoiced phones. Therefore, the observed F0 trajectory is usually discontinuous for whichever supra-segmental structure.

In standard HMM-based synthesis this problem is avoided by modelling directly the observed discontinuous log-F0 at a sub-segmental level by means of multi-space distributions (MSD) [2]. In an MSD, the log-F0 signal is assumed to be either a random variable sampled from a 1-dimensional distribution for voiced frames, or a 0-dimensional symbol for unvoiced ones. At synthesis time, the prior probability of these two spaces is used to classify each frame into voiced and unvoiced. A continuous F0 trajectory is then generated for each sub-section of voiced frames using the standard parameter generation algorithm [3]. In the original full HMM-based TTS [4] the predicted F0 had to be discontinuous because it was also used to control the pulse/noise switch excitation model. Nowadays, most HMM-based TTS system uses a more sophisticated excitation scheme in which the voicing does not depends only on the predicted F0 values. Those systems perform better when the voicing is controlled by the frequency-dependent soft-decision provided by the excitation parameters rather than by the frequency-independent hard-decision of the predicted discontinuous F0 [5]. Moreover, relieving F0 from any responsibility regarding the voicing allows treating it as a continuous signal, thus improving the intonation model [6, 7, 8, 9]

Another problem of the standard MSD model is that each voiced section is generated independently. Supra-segmental structures are ignored or at most considered only implicitly via the decision tree used to select the models. A proposed method to generate F0 using explicit supra-segmental information consists of obtaining the log-F0 contour that maximizes the weighted sum of log-likelihoods of several intonation models, each at a different linguistic level [10, 11]. This approach can produce a better intonation than a standard state-based MSD model [10, 11, 12]. The supra-segmental model consists of distributions of a fixed-order parametrisation of the log-F0 contour at that level. However, some parametrizations are unstable when applied to discontinuous signals. The standard way to deal with this problem is to interpolate log-F0 [13], usually by a linear or a spline function with a window of one or two frames before and after the unvoiced gap. This causes two new problems. First, F0 values close to the unvoiced regions are often unreliable. Therefore, the values interpolated from them are unreliable too [7]. Second, interpolated values rarely follow the 'natural' contour of the data. As a result, they introduce a distortion in the coefficients which might affect the model [14]. It is possible to avoid this by computing the parametrisation only over continuous F0 sections [15, 11]. However, this makes building statistical models harder, because the meaning of the coefficients depend completely of the underlying phonetic structure. For example, two phonetically different syllables, e.g., 'big' and 'pick', pronounced with the same intonation might have different coefficients because the boundaries of their voiced sections are different.

This paper investigates two approaches to obtain parametrisation coefficients over whole linguistic units without interpolating F0. The rest of the paper is organized as follows. Section 2 reviews the parametric F0 approach and its similarities to other continuous F0 models. Section 3 introduces the two proposed method to avoid interpolation. Section 4 shows the result of a subjective experiment. A possible explanation for these results is discussed in section 5. Finally, conclusions are drawn in 6.

[*]Kate Knill is currently at Cambridge University.

2. PARAMETRIC F0 MODEL

In statistical intonation methods, the F0 or log-F0 signal is considered to be a random variable with a distribution $P(\boldsymbol{x}|T,\boldsymbol{\lambda})$ that depends on the text T. The mapping between T and $\boldsymbol{x}$ is defined by the intonation model $\boldsymbol{\lambda}$. During training, supervised learning is used to obtain the parameters of $\boldsymbol{\lambda}$. During synthesis, the parameters of $\boldsymbol{\lambda}$ are combined to obtain the appropriate $P(\boldsymbol{x}|T,\boldsymbol{\lambda})$ and the $\boldsymbol{x}$ that maximizes it is generated.

In a parametric F0 model, instead of modelling $\boldsymbol{x}$ directly, a parametrisation of the segments of $\boldsymbol{x}$ chunked at the considered linguistic level is modelled. When the parametrization is linear

$$\boldsymbol{x} = \boldsymbol{N}\boldsymbol{c} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{N}$ is a block diagonal matrix formed by the concatenation of the inverse parametrization matrices of all the units at the considered level. The error term $\boldsymbol{\epsilon}$ is assumed to follow a Gaussian distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$ with $\boldsymbol{V}$ a diagonal matrix. If $\boldsymbol{\lambda}$ consists of Gaussians

$$P(\boldsymbol{x}|\boldsymbol{\lambda}) = \int P(\boldsymbol{x}|\boldsymbol{c},\boldsymbol{\lambda})P(\boldsymbol{c}|\boldsymbol{\lambda})\mathrm{d}\boldsymbol{c} = \mathcal{N}(\boldsymbol{x};\hat{\boldsymbol{x}},\boldsymbol{U}) \tag{2}$$

with

$$\hat{\boldsymbol{x}} = \boldsymbol{N}\hat{\boldsymbol{c}} \tag{3}$$

$$\boldsymbol{U} = \boldsymbol{V} + \boldsymbol{N}\boldsymbol{P}\boldsymbol{N}^\top \tag{4}$$

where $\hat{\boldsymbol{c}}$ and $\boldsymbol{P}$ are the mean value and covariance of the trajectory of coefficients $\boldsymbol{c}$. If the model consists only of static coefficients $\hat{\boldsymbol{c}} = \boldsymbol{\mu}$. In that case the continuity of F0 at the joints between units is not guaranteed. To avoid this, the model includes a set of concatenation coefficients $\boldsymbol{\Delta c}$ such as the delta of the average log-F0 of the unit or the gradient of the log-F0 at the unit boundaries [16]. With these extra coefficients the observation vector becomes $\boldsymbol{o} = [\boldsymbol{c}^\top, \boldsymbol{\Delta c}^\top]^\top$. If $\boldsymbol{o} = \boldsymbol{M}\boldsymbol{c}$, the generation of $\boldsymbol{c}$ follows the same equations as the standard parameter generation algorithm[3]

$$\boldsymbol{P} = (\boldsymbol{M}^\top \boldsymbol{\Sigma}'\boldsymbol{M})^{-1} \tag{5}$$

$$\hat{\boldsymbol{c}} = \boldsymbol{P}\boldsymbol{M}^\top \boldsymbol{\Sigma}'^{-1}\boldsymbol{\mu} \tag{6}$$

where $\boldsymbol{\mu}'$ and $\boldsymbol{\Sigma}'$ are the mean and covariance of $\boldsymbol{o}$.

In the simplest case, the covariance of the error term $\boldsymbol{V}$ consists of the alternation of two global values, σ_o for frames with an observed F0 and σ_m for the missing ones. A more sophisticated model could define σ_o to be context dependent or a function of some other signal, e.g. the band aperiodicity.

2.1. Similarities to other continuous F0 models

At frame-level, the log-F0 distributions produced by this model are Gassian mixtures with two components, one for the observed frames and another for the missing ones. This is similar to the continuous F0 model proposed in [6], specially if σ_m is global and with a sufficiently large value so that its actual mean does not matter. The main difference between that model and the one defined by Eq. (1) is that here the means are shared. Therefore, the mean trajectory is always continuous and smooth, at least within the boundaries defined by the units of the considered linguistic level.

Another continuous F0 supra-segmental model that treats unvoiced regions as missing data was proposed in [7]. The main difference with that approach is that it models supra-segmental intonation with a 5-states HMM instead of by parametrizing it.

3. INITIALIZING THE MODEL

To train a parametric F0 model, first the static coefficients $\boldsymbol{c}_s$ of each unit are computed. Next, a set of concatenation coefficients $\Delta\boldsymbol{c}_s$ are obtained to form the observation vectors $\boldsymbol{o}_s$. The observation vectors are then clustered with a decision tree and finally, the model parameters can be refined by retraining the model as a trajectory model using, for example, a minimum-generation error criterion [16].

The main remaining problem of this approach is how to compute the observation vectors. Ideally, they should depend only on the observed F0. However, for some parametrizations such as the discrete cosine transform (DCT), coefficients obtained only from observed F0 values might be meaningless, because the transform is poorly conditioned for discontinuous signals[17]. In [10], this problem was circumvent by interpolating log-F0 with a local spline function. However, the interpolated values might introduce a distortion in the extracted coefficients which bias the model [14].

A similar problem was found in the computation of cepstral envelopes from discrete frequency points. Two approaches proposed to solve that problem were a) using a regularization term [18] and b) using a factor analysis (FA) approach in which the cepstral/DCT coefficients are considered as hidden variables. [19]. The next subsections describe how these approaches could be applied to training a supra-segmental parametric F0 model.

3.1. Regularization approach

Usually, the coefficients of a linear parametrization are computed using a least-square criterion. For one single unit s, the least-square solution of the model of eq. (1) is

$$\hat{\boldsymbol{c}}_s = (\boldsymbol{N}_s^\top \boldsymbol{W}_s \boldsymbol{N}_s)^{-1}\boldsymbol{N}_s^\top \boldsymbol{W}_s \boldsymbol{x}_s \tag{7}$$

where $\boldsymbol{W}_s$ is a diagonal weighting matrix. When $\boldsymbol{x}_s$ is continuous and/or no element of $\boldsymbol{W}_s$ is too small, eq. (7) can be solved. Otherwise $\hat{\boldsymbol{c}}_s$ might be unstable. To avoid this, a smoothness constraint R can be added to the least-squared error function

$$F(\boldsymbol{c}_s) = (\boldsymbol{x}_s - \boldsymbol{N}_s\boldsymbol{c}_s)^\top(\boldsymbol{x}_s - \boldsymbol{N}_s\boldsymbol{x}_s) + \rho R(\boldsymbol{N}_s\boldsymbol{c}_s)) \tag{8}$$

Applied to F0, the regularization proposed in [18] is

$$R(N,\boldsymbol{c}) = \int_0^{D_s} (\frac{\mathrm{d}N(\boldsymbol{c},t)}{\mathrm{d}t})^2 \mathrm{d}t \tag{9}$$

where D_s is the length in frames of s, and N is the continuous function from where matrix $\boldsymbol{N}_s$ is derived. For a DCT of order p

$$N(\boldsymbol{c},t) = \sqrt{2}c_0 + 2\sum_{i=1}^{p-1} c_i \cos(\pi \frac{(t+0.5)}{D} i) \tag{10}$$

Rewriting eq. (9) as

$$R(\boldsymbol{N}\boldsymbol{c}) = \boldsymbol{c}^\top \boldsymbol{R}\boldsymbol{c} \tag{11}$$

$$\boldsymbol{R} = \frac{2\pi^2}{D_s}\,\mathrm{diag}([0,1,2^2,\cdots,(p-1)^2) \tag{12}$$

the value that maximizes eq. (8) is

$$\hat{\boldsymbol{c}}_s = (\boldsymbol{N}_s^\top \boldsymbol{W}_s \boldsymbol{N}_s + \rho \boldsymbol{R}_{D_s})^{-1}\boldsymbol{N}_s^\top \boldsymbol{W}_s \boldsymbol{x}_s \tag{13}$$

The terms of $\boldsymbol{W}$ can be simply 1 for observed values and 0 for unobserved ones. The value of ρ could depend on the number of seen frames or just be constant, e.g. 10^{-4}.

Once obtained the DCT coefficients of each unit, the concatenation coefficients $\boldsymbol{\Delta c}$ are computed from their linear combination. Finally, the full observation vector is clustered together by a decision tree.

3.2. Factor analysis approach

The second method is based on the Joint Extraction and Modeling Approach (JEMA) [14] but using factor analysis to obtain the model for each cluster. In this approach, the parametrization coefficients are considered hidden variables. Therefore, instead of first extracting the DCT coefficients and then get their distribution, their distribution are computed directly using an expectation-maximization algorithm and the auxiliary function

$$Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = \sum_{\forall s \in j} \int P(\boldsymbol{c}|\boldsymbol{x}_{\mathtt{o}}, \boldsymbol{\lambda}) \log(P(\boldsymbol{c}, \boldsymbol{x}_{\mathtt{o}}|\hat{\boldsymbol{\lambda}}))d\boldsymbol{c} \tag{14}$$

where $\{s \in j\}$ is the subset of units associated to cluster j. When no $\boldsymbol{\Delta c}$ is considered, $\boldsymbol{P} = \boldsymbol{\Sigma}$ and $\hat{\boldsymbol{c}} = \boldsymbol{\mu}$. Thus, the model parameters and log-likelihood of each cluster j can be computed independently. Based on eq. (1) and assuming Gaussian distributions the maximization step yields the update equations

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{\forall s \in j} \bar{\boldsymbol{c}}_s}{S_j} \tag{15}$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{\sum_{\forall s \in j} \boldsymbol{\varphi}_s + \bar{\boldsymbol{c}}_s \bar{\boldsymbol{c}}_s^{\top}}{S_j} - \hat{\boldsymbol{\mu}}_j \hat{\boldsymbol{\mu}}_j^{\top} \tag{16}$$

where S_j is the number of units in j and $\boldsymbol{\varphi}_s$ and $\bar{\boldsymbol{c}}_s$ the covariance and mean of the posterior distribution $P(\boldsymbol{c}_s|\boldsymbol{x}_s, \boldsymbol{\lambda}))$, obtained in the expectation step as

$$\boldsymbol{\varphi}_s = (\boldsymbol{\Sigma}_j^{-1} + \frac{\boldsymbol{N}_{s,\mathtt{o}}^{\top} \boldsymbol{N}_{s,\mathtt{o}}}{\sigma_j})^{-1} \tag{17}$$

$$\bar{\boldsymbol{c}}_s = (\boldsymbol{\varphi}_s^{-1} + \frac{\boldsymbol{N}_{s,\mathtt{m}}^{\top} \boldsymbol{N}_{s,\mathtt{m}}}{\sigma_m})^{-1} (\boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j + \frac{\boldsymbol{N}_{s,\mathtt{o}}^{\top} \boldsymbol{x}_{s\mathtt{o}}}{\sigma_j}) \tag{18}$$

The sub-indices $\mathtt{o}$ and $\mathtt{m}$ refer to the observed and missing frames/rows in $\boldsymbol{x}_s$ and $\boldsymbol{N}_s$. The maximization update formula for σ_j is

$$\hat{\sigma}_j = \frac{\sum_{\forall s \in j} \mathrm{tr}(\boldsymbol{N}_{s,\mathtt{o}} \boldsymbol{\varphi}_s \boldsymbol{N}_{s,\mathtt{o}}^{\top}) + (\boldsymbol{x}_{s,\mathtt{o}} - \boldsymbol{N}_{s,\mathtt{o}} \bar{\boldsymbol{c}}_s)^{\top} (\boldsymbol{x}_{s,\mathtt{o}} - \boldsymbol{N}_{s,\mathtt{o}} \bar{\boldsymbol{c}}_s)}{\sum_{\forall s \in j} D_{s,\mathtt{o}}} \tag{19}$$

with $D_{s\mathtt{o}}$ the number of observed frames in $\boldsymbol{x}_s$. The expectation-maximization steps are then iterated a fixed number of times or until the model converges. The error term for missing frames $\sigma_{\mathtt{m}}$ is an invariant global constant defined during the initialization, e.g., 10^4. To reduce computational cost $\boldsymbol{\Sigma}_j$ can be diagonalized.

To obtain the model structure a methodology based on JEMA [14] is applied. First, a root node model is computed using above equations over all the training units. Then, for each question the models associated to the 'yes' and 'no' subsets are computed. Finally, the question that produces the split with the largest increment of log-likelihood with respect to the parent model is selected.This process is repeated for each new node until an stopping criterion is reached, e.g. maximum description legth (MDL) [20].

When the number of training units or the number of questions is large, as usually is the case, this process becomes extremely expensive. To simplify it, the questions can be pre-filtered by assuming that the optimum question in the maximum-likelihood sense will be among the top 30-50 questions in the least-square sense, which are easier to obtain. First, the that minimizes the average least-square error over all the units of a node j is computed as $\tilde{\boldsymbol{\mu}}_j = \boldsymbol{a}_j^{-1} \boldsymbol{b}_j$ with

$$\boldsymbol{a}_j = \sum_{\forall s \in j} \boldsymbol{\alpha}_s^{\circ} = \sum_{\forall s \in j} \boldsymbol{N}_{s,\mathtt{o}}^{\top} \boldsymbol{N}_{s,\mathtt{o}} \tag{20}$$

$$\boldsymbol{b}_j = \sum_{\forall s \in j} \boldsymbol{\beta}_s^{\circ} = \sum_{\forall s \in j} \boldsymbol{N}_{s,\mathtt{o}} \boldsymbol{x}_{s,\mathtt{o}} \tag{21}$$

The total error for $\tilde{\boldsymbol{\mu}}_j$ is

$$e(\tilde{\boldsymbol{\mu}})_j = g_j - \boldsymbol{b}_j^{\top} \boldsymbol{a}_j^{-1} \boldsymbol{b}_j \tag{22}$$

where

$$g_j = \sum_{\forall s \in j} \boldsymbol{x}_{s,\mathtt{o}}^{\top} \boldsymbol{x}_{s,\mathtt{o}} \tag{23}$$

does not change by splitting j. Therefore, the error reduction obtained by a split is

$$\Delta e_{p,y,n} = \boldsymbol{b}_y^{\top} \boldsymbol{a}_y^{-1} \boldsymbol{b}_y + \boldsymbol{b}_n^{\top} \boldsymbol{a}_n^{-1} \boldsymbol{b}_n - \boldsymbol{b}_p^{\top} \boldsymbol{a}_p^{-1} \boldsymbol{b}_p \tag{24}$$

with y,n and p indices indicating 'yes','no' and 'parent'. Eq. 24 can be calculated very efficiently by pre-computing $\boldsymbol{\alpha}_s$ and $\boldsymbol{\beta}_s$ for each unit.

Running 4-5 iterations of the EM algorithm for the children of 30-50 questions is reasonably quick but computing the log-likelihood for all of them is still quite expensive. A second filtering criterion that can be used is the improvement produced by the split over the auxiliary function. With these two question filters, the number of final splits for which the log-likelihood has to be computed can be reduced from several thousands to just 3 or 5.

Any EM algorithm requires an initial model. For the root node a reasonable one can be obtained from the least-square solution

$$\boldsymbol{\mu}_0 = \tilde{\boldsymbol{\mu}}_0 = \boldsymbol{a}_0^{-1} \boldsymbol{b}_0 \tag{25}$$

$$\boldsymbol{\Sigma}_0 = \boldsymbol{I} \tag{26}$$

$$\sigma_j = \frac{e(\tilde{\boldsymbol{\mu}})_0}{\sum_{\forall s} D_{s,\mathtt{o}}} \tag{27}$$

with $\boldsymbol{I}$ an identity matrix. For the rest of splits, $\boldsymbol{\mu}_j$ and σ_j can be initialized also with the least-squared solution and Σ_j can be copied from the parent node.

Once the static model is trained, the model of the concatenation coefficients is computed from the posterior distributions of the static coefficients of each unit

$$P(\boldsymbol{c}_s|\boldsymbol{x}_{s,\mathtt{o}}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{c}; \bar{\boldsymbol{c}}_s, \boldsymbol{\varphi}_s) \tag{28}$$

For each unit, the distribution of its concatenation coefficients is obtained as the linear combination of the posterior distributions of the static coefficients of current and neighbour units. The concatenation model is obtained by clustering these distributions, either forcing the decision tree of the static coefficients or with an independent decision tree.

4. EXPERIMENTS

4.1. Compared Models

To test the different approaches 4 syllable-level DCT models were created:

•'interp' baseline model: the static coefficients were obtained from a spline interpolation of 'reliable' F0 values [16] and the full observation vector was clustered using HTS [21].

•'reg' model: the static coefficients were obtained from the observed sections of F0 using the regularization approach and the full observation vector was clustered using HTS.
•'FA-sta' model: the model for the static coefficients was trained using the factor analysis approach. The decision tree of the static coefficients was then imposed over the concatenation coefficients.
•'FA-rcl' model: same static model as 'FA-sta' but the distributions of the concatenation coefficients were re-clustered independently using HTS.
The static coefficients where a 5-order DCT. The concatenation coefficients consisted of the delta of the 0-th DCT coefficient and the log-F0 gradient at the beginning and end of the syllable [16]. The number of final leaves for 'interp', 'reg' and 'FA-sta' was 5094, 4768 and 4111 respectively. The number of concatenation coefficient leaves of 'FA-rcl' was 460.

The question pre-selection scheme used for the FA training was as follows. First, 50 questions were pre-selected based on the least-squared criterion. From them, 5 were selected based on ΔQ and finally the one with maximum-likelihood improvement was chosen to split the node. A 5-fold cross-validation was applied in the least-square pre-selection. Only those questions that were able to produce a valid split over all 5 folds were considered.

4.2. Training data

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 was 65129. The database was automatically annotated both phonetically and syntactically. In addition to the usual phonetic and syntactic questions, questions about the duration of the syllable, its head and its coda were added [22]. To get this features, the durations for the labels were obtained from a Viterbi forced alignment of the database with the same frame-level HSMM model used to generate the duration, spectrum and aperiodicity during synthesis.

4.3. Experimental conditions and results

A set of subjective preferences test were conducted, each consisting of 162 paired-stimuli (81 unique sentences each presented as AB and BA) which were judged by 5 different subjects. Subjects could choose a 'none' option if they considered both stimuli equal, but were encouraged not to use it too often. The evaluation was crowd-sourced to Amazon Mechanical Turk via CrowdFlower. All the tests were run simultaneously. The results were filtered to reduce the effect of possible spammers [23]. After the anti-spam filters, the average number of individual subjects and average number of judgements per test was 178 and 730 respectively.

Table 4.3 shows the results. The model trained from the interpolated F0 is clearly the worst one. This seems to confirm the hypothesis about the distortion introduced by the F0 interpolation. The regularization model seems to outperform both factor analysis ones which are equivalent.

5. DISCUSSION

The results regarding the interpolated model were expected. However, the better performance of the regularized model w.r.t. the FA ones was surprising because the FA approach is based on a proper statistical framework whereas the regularization one is based on a more or less heuristically defined penalty term. The main difference between the equations to estimate $\boldsymbol{c}$ in the regularization and the FA method, (8) and (18) respectively, is the prior term $\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j$ added to the numerator of the FA one. This term guarantees that there is a posterior distributions even for units with no observed F0 value. The ratio between $|\boldsymbol{\Sigma}_j|$ and σ_j gives the relative importance assigned to that prior versus the observation $\boldsymbol{x}_{s,\mathrm{o}}$. Assuming and initial value of $\sigma_j \ll |\Sigma_j|$, that is, most of the weight is on the observation

$$\boldsymbol{\varphi}_s \simeq \sigma_j(\boldsymbol{N}_{s,\mathrm{o}}\boldsymbol{N}_{s,\mathrm{o}}^\top)^{-1} \quad (29)$$

$$\mathrm{tr}(\boldsymbol{N}_{s,\mathrm{o}}\boldsymbol{\varphi}_s\boldsymbol{N}_{s,\mathrm{o}}^\top) \simeq D_{s,\mathrm{o}}\sigma_j \quad (30)$$

and the update equation of σ_j can be approximated by

$$\hat{\sigma}_j \simeq \sigma_j + \frac{\sum_{\forall s\in j}(\boldsymbol{x}_{s,\mathrm{o}} - \boldsymbol{N}_{s,\mathrm{o}}\bar{\boldsymbol{c}}_s)^\top(\boldsymbol{x}_{s,\mathrm{o}} - \boldsymbol{N}_{s,\mathrm{o}}\bar{\boldsymbol{c}}_s)}{\sum_{\forall s\in j} D_{s,\mathrm{o}}} \quad (31)$$

The second summation term in eq. (31) can be positive or zero. Therefore, each iteration might tend to increase σ_j until $\sigma_j \gg \boldsymbol{\Sigma}_j$. Then $\boldsymbol{\varphi}_s \simeq \boldsymbol{\Sigma}_j$ and $\bar{\boldsymbol{c}}_s \simeq \boldsymbol{\mu}_j$. At this point the model parameters cannot change any more, i.e., the model has converged. Moreover, $P(\boldsymbol{c}_s|\boldsymbol{x}_s,\boldsymbol{\lambda}) \simeq P(\boldsymbol{c}_s|\boldsymbol{\lambda})$. Therefore $\boldsymbol{\Delta c}$ coefficients computed from those posteriors are roughly the same as those that might be computed from the priors. This implies that, during synthesis, the distributions of $\boldsymbol{\Delta c}$ already match the linear combination of the distributions of $\boldsymbol{c}$ and thus, they do not contribute with any new information/constraint. This might explain why the structure of the concatenation coefficients has no effect on the quality of the generated model. It might also explain the preference for the regularization approach, in which the concatenation coefficients do add extra information.

interp	reg	FA-sta	Fa-rcl	None	p-score
12.4	81.8	-	-	5.8	$< 10^{-3}$
18.2	-	74.9	-	6.9	$< 10^{-3}$
19.3	-	-	74.6	6.1	$< 10^{-3}$
-	50.5	31.9	-	17.6	$< 10^{-3}$
-	51.6	-	33.7	14.8	$< 10^{-3}$
-	-	37.7	37.3	25.1	0.46

Table 1. Preference scores for different syllable-level DCT-F0 models

6. CONCLUSIONS AND FUTURE WORK

Two methods to initialize a parametric F0 models using only observed F0 values have been proposed, one based on regularization and the other on factor analysis. Subjective experiments have shown a significant preference for both methods w.r.t. a baseline using standard F0 interpolation. This confirms the hypothesis that the interpolated values distort the model. Despite its simplicity the regularization method outperforms the factor analysis one. This might be due to the fact that the concatenation coefficients of the factor analysis models provide little or no information.

The next step is to fully re-train the model as a trajectory, either by itself or together with other level models within a product-of-experts framework [24].

7. REFERENCES

[1] G. Kochanski and C. Shih, "Stem-ml: Language-independent prosody description," in *Proc. ICSLP*, 2000, pp. 239–242.

[2] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. ICASSP*, 1999, pp. 229–232.

[3] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, 1995, pp. 660–663.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[5] J. Latorre, M. Gales, S. Buchholz, K. Knill, M. Tamura, Y. Ohtani, and M. Akamine, "Continuous F0 in the source-excitation generation for HMM-based TTS:Do we need voiced/unvoiced classification?" in *Proc. ICASSP*, 2010, pp. 4724–4727.

[6] K. Yu and S. Young, "Continuous F0 modelling for HMM based statistical parametric speech synthesis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 99, pp. 1071–1079, 2011.

[7] T. Koriyama, T. Nose, and T. Kobayashi, "Discontinuous observation HMM for prosodic-event-based F0 generation," in *Proc Interspeech*, 2012.

[8] W. Wang, M. Wen, D. Saito, K. Hirose, and N. Minematsu, "Improved generation of prosodic features in HMM-based mandardin speech synthesis," in *Proc. 7th Speech Synthesis Workshop*, 2010, pp. 359–364.

[9] Q. Zhang, F. Soon, Y. Qian, Z. Yan, J. Pan, and Y. Yan, "Improved modeling for F0 generation and V/U decition in HMM-based TTS," in *Proc. ICASSP*, 2010, pp. 4606–4609.

[10] J. Latorre and M. Akamine, "Multilevel parametric-base f0 model for speech synthesis," in *Proc. Interspeech*, 2008, pp. 2274–2277.

[11] Y. Qian, Z. Wu, and F. Soong, "Improved prosody generation by maximizing joint likelihood of state and longer units," in *Proc. ICASSP*, 2009, pp. 3781–3784.

[12] N. Obin, A. Lacheret, and X. Rodet, "Stylization and trajectory modelling of short and long term speech prosody variations," in *Proc. Intesrspeech*, 2011, pp. 2029–2032.

[13] H. Pfitzinger, H. Mixdorff, and J. Schwarz, "Comparison of Fujisaki-model extractors and F0 stylizers," in *Proc. Interspeech*, 2009, pp. 2455–2458.

[14] P. Aguero, J. Tulli, and A. Bonafonte, "A study of jema for intonation modeling," in *Proc ICASSP*, 2008, pp. 4625–4628.

[15] J. Teutenberg, C. Watson, and P. Riddle, "Modelling and synthesising F0 contours with the discrete cosine transform," in *Proc. ICASSP*, 2008, pp. 3973–3976.

[16] J. Latorre, M. Gales, and H. Zen, "Training a parametric-based logf0 model with the minimum generation error criterion," in *Proc. Interspeech*, 2010, pp. 2174–2177.

[17] C. Lawson and R. Hanson, *Solving Least-Squares Problems*. Prentice Hall, 1974.

[18] O. Cappe, J. Laroche, and E. Moulines, "Regularized estimation of cesptrum envelope from discrete frequency points," in *Proc. ASSP*, 1995, pp. 213–216.

[19] T. Toda and K. Tokuda, "Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory hmm," in *Proc ICASSP*, 2008, pp. 3925–3928.

[20] J. Rissanen, *Stochastic complexity in stochastic inquiry*. World Scientific Publishing Company, 1980.

[21] K. Tokuda, H. Zen, S. Sako, T. Yoshimura, J. Yamagishi, M. Tamura, and T. Masuko, "The HMM-based speech synthesis software toolkit," http://hts.sp.nitech.ac.jp/.

[22] J. Latorre, S. Buchholz, and A. Akamine, "Usages of an external duration model for HMM-based speech synthesis," in *Proc. Speech Prosody 2010*, 2010.

[23] S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating," in *Proc. Interspeech*, 2011, pp. 3053–3056.

[24] H. Zen, M. Gales, Y. Nankaku, and T. Tokuda, "Product of experts for statistical parametric speech synthesis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 3, pp. 794–805, 2012.