# PROBABLISTIC MODELLING OF F0 IN UNVOICED REGIONS IN HMM BASED SPEECH SYNTHESIS

*K. Yu, T. Toda*, M. Gasic, S. Keizer, F. Mairesse, B. Thomson and S. Young*

Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, UK
Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630-0192,Japan
Email: {ky219, mg436, sk561, farm2, brmt2, sjy}@eng.cam.ac.uk, tomoki@is.naist.jp

## ABSTRACT

HMM based synthesis has attracted great interest due to its compact and flexible modelling of spectral and prosodic parameters. In this approach, short term spectra, fundamental frequency (F0) and duration are simultaneously modelled by multi-stream HMMs. However, since F0 values in unvoiced regions are normally considered as undefined, it is difficult to use standard HMMs for F0 modelling. The currently preferred solution to this is to use a multi-space distribution HMM (MSDHMM) in which discrete distributions are used for modelling the voiced/unvoiced decision and continuous Gaussian distributions are used for modelling the F0 values within the voiced regions. However, the assumption of undefined unvoiced F0 regions and the special structure of the MSDHMM lead to limitations in the accurate modelling of F0 patterns. In this paper an alternative is explored whereby unvoiced F0 values are assumed to exist and are modelled within the standard HMM framework using a globally tied distribution (GTD). Subjective evaluations show that these regular HMMs with GTD can produce significant improvements in the naturalness of the synthesised speech compared to the MSDHMM, and furthermore, the method is insensitive to the exact method used for unvoiced F0 generation.

***Index Terms***— HMM based synthesis, F0 modelling

## 1. INTRODUCTION

As an alternative to the traditional unit concatenation approach, HMM-based speech synthesis has attracted considerable interest recently due to its compact and flexible representation of voice characteristics [1]. Based on the source-filter model of speech production, spectral features, excitation features, fundamental frequency (F0) and duration are modelled as separate streams within a set of context-dependent phone-level HMMs [1]. These phone models are trained from parameters extracted (eg using STRAIGHT[2]) from a corpus of utterances spoken by a single speaker. In the synthesis stage, the source text is converted to a phoneme sequence and the corresponding HMMs are concatenated. The resulting composite HMM is then used to generate a sequence of parameters [3] which are converted to a waveform using an appropriate synthesis filter [4].

In HMM-based synthesis, the modelling of fundamental frequency (F0) is difficult due to the discontinuity of F0 values across voiced and unvoiced regions. During voiced speech, the periodic airflow modulation generated at the glottis serves as the excitation for the vocal tract and since there exists strong periodicity, F0 values can be effectively estimated from the waveform [5]. However, unvoiced speech is produced when the airflow is forced through a vocal-tract constriction with sufficient velocity to generate significant turbulence. The long term spectrum of turbulent airflow tends to be a weak function of frequency [6] and hence F0 values during unvoiced regions require special treatment. Normally, these F0 values are assumed to be undefined but it is then not possible to use a simple continuous distribution to model F0.

The multi-space distribution HMM (MSDHMM) provides a solution to this problem by using a combination of discrete and continuous distributions [7] and it is now the default modelling approach in state-of-the-art HMM synthesis systems. However, although good performance can be achieved using MSDHMMs, this type of mixed distribution F0 modelling has some issues arising from the discontinuities at the boundaries of unvoiced regions and the need to keep the discrete and continuous density regions distinct. Furthermore, the use of MSDHMMs makes it more difficult to exploit standard techniques for HMM modelling, such as adaptation, which cannot be readily applied to the mixed discrete/continuous F0 distributions.

In this paper, a simpler alternative to the MSDHMM is investigated in which it is assumed that F0 values do exist and are observable in both voiced and unvoiced regions. Hence, standard HMMs can be used. Dynamic features can then be calculated for all frames and only one stream is required for F0 modelling. This assumption has been used in stress classification [8] and in intonation contour modelling [9], but it has not been investigated for HMM based synthesis so far. Note that this assumption does not imply that F0 is simply interpolated across unvoiced regions using some form of curve fitting. Instead, it is assumed that the F0 values observed in unvoiced regions are drawn from a different distribution. In practice, these observations can be selected from the F0 candidates generated by the pitch tracker during the feature extraction stage or sampled from some pre-defined distribution. Thus, the F0 stream is modelled by a Gaussian mixture with two components, corresponding to voiced and unvoiced F0, respectively. Due to the nature of unvoiced speech, mixtures corresponding to unvoiced F0 values are globally tied. Hence, the approach is referred to as a *HMM with a globally tied distribution (GTD)*. As will be shown below, the HMM-GTD model is not only simpler but subjective listening tests also show that it can significantly improve the naturalness of synthesised speech.

The rest of the paper is arranged as follows. Section 2 discusses F0 modelling in the MSDHMM in some detail. The HMM-GTD model is then described in section 3. Section 4 presents the experimental results, followed by the conclusions in section 5.

---

The $2^{nd}$ author did the work while visiting Cambridge University.

This research was partly funded by the UK EPSRC under grant agreement EP/F013930/1 and by the EU FP7 Programme under grant agreement 216594 (CLASSIC project: www.classic-project.org).

[1] Further information such as aperiodic energy may also be modelled using additional streams within the same HMM framework.

## 2. F0 MODELLING IN HMM BASED SYNTHESIS

As indicated in section 1, a common assumption is that F0 is a continuous value in voiced (v) regions and it is undefined in unvoiced (uv) regions. This implies that for HMM-based synthesis, a discrete v/uv indicator is required at each frame and this places some constraints on the associated HMMs.

Firstly, accurate F0 modelling within a HMM framework requires that the static F0 values are augmented with dynamic parameters consisting of the $1^{st}$ and $2^{nd}$ order derivatives. For frames at the boundary between voiced and unvoiced regions, this causes problems because the dynamic features are not defined at discontinuities. Although methods of using a single stream have been considered (eg. [10]), the preferred solution to this problem is to regard these undefined values as being unvoiced [11]. This means that the boundaries differ for static and dynamic parameters and consequently separate streams have to be used for each of the F0 features. The effect of this is that the correlation modelling between the static and dynamic F0 parameters is weakened. During training there will be a greater tendency to overfit the data, and during synthesis, the reduced constraints on the temporal correlation may degrade the accuracy of the generated F0 trajectories.

Secondly, in order to simultaneously model the discrete v/uv decision and the continuous F0 trajectory variables, multi-space distribution HMMs (MSDHMM) are commonly used [7]. The state output distribution in an MSDHMM is

$$b_\theta(o) = \begin{cases} c_v \mathcal{N}(o; \mu_\theta, \sigma_\theta) & \text{o} \in \text{voiced region} \\ c_{uv} & \text{o} \in \text{unvoiced region} \end{cases} \quad (1)$$

where $o$ is the observation at state $\theta$, $c_v$ and $c_{uv}$ ($c_v + c_{uv} = 1$) are the probabilities of voiced and unvoiced regions, and $\mu_\theta$ and $\sigma_\theta$ are the means and variances of the Gaussian distribution of F0 in the voiced regions. During synthesis, each HMM state is first classified as voiced or unvoiced according to whether $c_v$ of the static stream is greater than 0.5. Then, for unvoiced states, white noise is used as the excitation while for voiced states, F0 values are generated as the excitation parameters.

This multi-space HMM framework results in some inherent limitations. Since $b_\theta(o)$ represents a continuous density in voiced regions and a discrete probability mass in unvoiced regions, each observation can only be either voiced or unvoiced, but not both at the same time. Consequently, during the forward-backward calculation for any F0 stream in training, the state posterior occupancy will always be wholly assigned to one of the two components depending on the voicing condition of the observation. This hard assignment limits the ability of the unvoiced component to learn from voiced data and vice versa, and it prevents any possibility of using a soft assignment to reduce the effect of F0 estimation errors.

A further problem is that the use of separate F0 streams introduces redundant mixture weights ($c_{uv}$). Hence, the number of free parameters is unnecessarily increased and this may have a significant effect on the state clustering where the minimum description length (MDL) criterion [12] is normally used. Using MDL, the description length of a model $\mathcal{M}_i$ given a data set $\mathcal{D}$ is defined as:

$$l(\mathcal{M}_i) \quad = \quad -\log p(\mathcal{D}|\mathcal{M}_i^{\mathtt{ML}}) + \lambda \frac{\alpha_i}{2} \log N_\mathcal{D} + K. \quad (2)$$

where $\mathcal{M}_i^{\mathtt{ML}}$ is the maximum likelihood (ML) estimate of $\mathcal{M}_i$ given the data, $\alpha_i$ is the number of free parameters in $\mathcal{M}_i$, $N_\mathcal{D}$ is the number of data points in $\mathcal{D}$[2], $K$ is a constant usually unchanged during clustering, and $\lambda$ is a factor controlling the weight of the model complexity part. From equation (2), the first term is the negative log likelihood, while the second term reflects the model complexity. The MDL criterion selects the model with the minimum description length for the given data. Hence, an increased model complexity will result in a smaller number of states following the decision-tree-based clustering. Thus, the redundant parameters in the MSDHMM may result in an underestimation of the cluster states. Furthermore, since the mixture weights for static and dynamic F0 streams are independent of each other, they may be very different in the final HMM after re-estimation. Thus, the forced use of separate streams may lead to a voicing classification which is inconsistent across streams.

[2]In HMM state clustering, $N_\mathcal{D}$ will be replaced by the total posterior occupancies of $\mathcal{M}_i$.

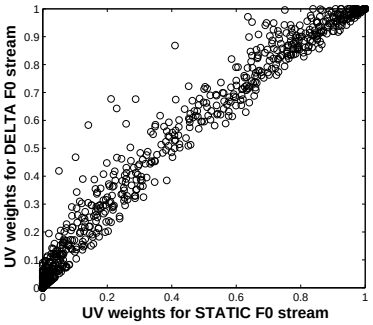


**Fig. 1**. Distribution of unvoiced mixture weights ($c_{uv}$)

Figure 1 shows the distribution of mixture weights for the static and delta F0 streams in an MSDHMM model trained with $\lambda = 0.6$ on the `slt` voice in the ARCTIC corpus as described in Section 4. The inconsistency of the weights for unvoiced speech between static and delta F0 streams can be clearly observed. It is also interesting to note that the delta weights are never less than the static weights. As explained above, this is a consequence of treating undefined delta coefficients as unvoiced thereby overestimating $c_{uv}$ in the delta stream [11]. Further evidence of this bias comes from our own experiments using the MSDHMM where we found that about 7.4% of the voiced states would have been classified as unvoiced if the delta stream weights had been used in place of the static stream weights.

## 3. HMM WITH GLOBALLY TIED DISTRIBUTION (GTD)

The previous section highlighted some of the problems encountered in HMM-based synthesis when F0 is assumed to be undefined in unvoiced regions. In the model described in this section, an alternative assumption is made whereby F0 values are assumed to exist in unvoiced regions but have markedly different statistical properties compared to their values in voiced regions. One motivation of this assumption comes from considering the synthesis stage. When generating the excitation for the synthesis filter, random noise is used for unvoiced regions. As with periodic excitations, this noise can be defined in the frequency domain, it is just the statistical characteristics that differ.

To implement a model where F0 is considered to exist in both voiced and unvoiced regions, two issues need to be addressed: how to obtain F0 observations within unvoiced regions and how to model the statistical difference between voiced and unvoiced regions.

A simple approach to obtaining F0 observations in unvoiced regions is to make use of the pitch tracker used in F0 extraction (here the STRAIGHT system is used [2]). In many pitch trackers, multiple F0 candidates are generated for each frame regardless of whether it is voiced or unvoiced. For training in unvoiced regions,

the first F0 candidate output by the pitch tracker may be selected as the F0 observation. This will be referred to as the *1-Best* selection. Alternatively, interpolation may be applied within the unvoiced regions and then the F0 candidate closest to the interpolated F0 track at each frame is selected as the observation. This will be referred to as *interpolation-based* selection. Finally, following previous stress classification work [8], pseudo-F0 observations could be used whereby log F0 values are sampled from a pre-defined Gaussian distribution with large variance.

Given the assumption that F0 observations are defined at every frame, dynamic F0 features can be calculated straightforwardly without considering any v/uv boundary effects. Consequently, static, delta and delta-delta F0 features can be modelled in a single stream.

To model the statistical difference between voiced and unvoiced F0 values, a two component GMM is used for the F0 stream. It is further assumed that the statistical property of unvoiced F0 observations are independent of state contexts. Hence, a *globally tied distribution (GTD)* is used to model all unvoiced F0 values. The likelihood of observing the F0 feature $o$ at state $\theta$ is therefore given by

$$b_\theta(o) \quad = \quad c_{uv}\mathcal{N}(o;\mu_{uv},\sigma_{uv}) + c_v\mathcal{N}(o;\mu_\theta,\sigma_\theta) \qquad (3)$$

where $c_{uv}+c_v=1$ are the weights for unvoiced and voiced components respectively, $\mu_{uv}$ and $\sigma_{uv}$ are the globally shared parameters of the unvoiced component, and $\mu_\theta$ and $\sigma_\theta$ are the state-dependent mean and variance parameters for the voiced components.

To initialise a model set prior to training, the global unvoiced Gaussian component can be either fixed to some pre-defined Gaussian distribution or it can be estimated as the global distribution over all unvoiced F0 values found by the pitch tracker. The subsequent training process is identical to standard HMM training and both the state-based Gaussians and the globally tied distribution are updated via the forward-backward algorithm.

In the synthesis stage, the required voicing classification is based on the component weights as in the MSDHMM. The parameters of the voiced components are then used for generating F0 values for voiced regions.

By making the unvoiced F0 existence assumption, the problems in section 2 are effectively addressed. Since there is only one single F0 stream, there are no redundant component weights parameters. Therefore, there will be no inconsistency in voicing classification and when using the MDL criterion in state clustering, more clustered states will be generated for the same $\lambda$ allowing the system to model richer F0 variations. Furthermore, the use of a single stream for F0 will yield stronger temporal modelling and the consistent use of continuous densities across both voiced and unvoiced regions allows soft v/uv boundary alignment in training. The latter combined with the "background" GTD component may be expected to mitigate some F0 extraction errors leading to more robust estimation of the voiced components.

## 4. EXPERIMENTS

To evaluate the performance of the HMM-GTD based synthesis framework compared to the MSDHMM, subjective listening preference tests were performed. The training data was taken from the CMU ARCTIC speech database. Data from two speakers, a U.S. female English speaker `slt` and a Canadian male speaker `jmk`, were used. Each data set has the same 1132 phonetically balanced sentences and is about 0.95 hours in duration. Both systems were built using the HTS HMM-synthesis toolkit version 2.1 [13].

The static feature set comprised 24 Mel-Cepstral coefficients, logarithm of F0 and aperiodic energy components in five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6 and 6 to 8 KHz). All features were extracted using the STRAIGHT speech analysis system [2]. Spectral, F0 and aperiodic component features were modelled using separate HMM streams. For the MSDHMM, the F0 features were further separated into three streams to separately model static, delta, delta-delta F0 features [11]. In contrast, the HMM-GTD system used a single stream to model all F0 features. Within unvoiced regions, the first F0 candidates extracted by STRAIGHT were used as the F0 observations (1-Best selection).

During HMM training for both systems, the stream weight for the aperiodic component was set to zero. Hence, the forward-backward alignments depended only on the spectral and F0 features. Statistics for the aperiodic components were however collected and their parameters were updated in the normal way.

After training the spectral and F0 model components, a single Gaussian duration model was estimated for each state from Viterbi alignments of the training data. This model was then used in the synthesis stage to generate state and phoneme durations. Note that a separate state clustering process was applied to obtain the duration model parameters.

Before discussing the synthesis quality of the two systems, it is informative to examine the training process. Firstly, the effect on the HMM training process was analysed. With the same MDL factor $\lambda = 1.0$, after clustering, HMM-GTD yielded significantly more F0 states (2176 for `slt` and 3180 for `jmk`) than MSDHMM (1043 for `slt` and 1880 for `jmk`). This demonstrates the effect of the redundant parameters in the MSDHMM. To give a fairer comparison between the MSDHMM and the HMM-GTD, the MDL factor of the MSDHMM was tuned so that the resultant model had similar number of clustered states (2032 for `slt` and 3079 for `jmk`) as the HMM-GTD systems. It is worth noting that even when the numbers of states were similar, the likelihood of the F0 part given the HMM-GTD was much larger than than that given the MSDHMM. Even though the likelihoods are not strictly comparable due to the different nature of the discrete and continuous density distributions, the significant likelihood difference nevertheless suggests that HMM-GTD model is making more efficient use of its parameters. To further demonstrate this, one training sentence from the female `slt` corpus was re-synthesised using both the HMM-GTD and the MSDHMM with similar number of states. The corresponding F0 values were then extracted using STRAIGHT and compared to those of the original speech. The comparison is shown in figure 2. Although there is a difference in duration, it can be observed that the F0 trajectory of the HMM-GTD is more similar to the original speech than for the MSDHMM, especially at the end of the phrase. When listening to the speech, it is also obvious that both the original and the HMM-GTD synthesised speech had a distinct rise at the end, while the MSDHMM speech was flat. This indicates that the HMM-GTD was a better model to represent the F0 trajectory.

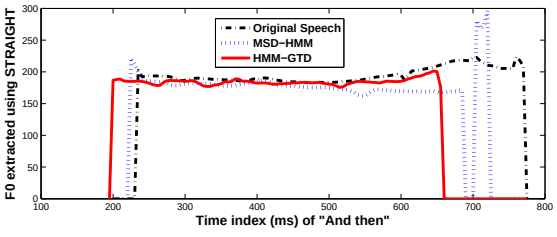


**Fig. 2**. Comparison of F0 trajectories

A subjective listening test was then performed to compare the

HMM-GTD and the MSDHMM with a similar number of F0 states. The listening test was a preference choice test on 36 sentences. Two wave files were synthesised for each sentence and each speaker voice using both the HMM-GTD and the MSDHMM. 12 sentences were then randomly selected to make up a testset for each listener, leading to 24 wave files pairs (12 for each voice). To reduce the variance introduced by forcing the user to make a choice, the 24 wave file pairs were duplicated and the order of the two systems were swapped. The final 48 samples were then shuffled and provided to the listeners. Each listener was asked to select the more natural example from each wave file pair. Altogether 21 listeners, 11 native and 10 non-native, participated in the listening test. The result is shown in figure 3.

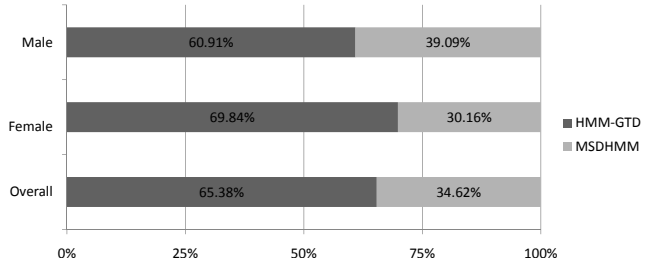


**Fig. 3**. Subjective comparison between the MSDHMM and the HMM-GTD based synthesisers for male and female speakers.

Statistical significance tests were performed assuming a binomial distribution of each choice. It was found that the HMM-GTD was significantly better than MSDHMM for both male and female speakers at a $p$ value of 0.01. It can also be observed that the listeners' preferences for the HMM-GTD generated speech were stronger for the female voice. This may be an artifact of the ARCTIC data because there is some background noise in the low fundamental frequency region for the female speaker and hence the GTD is more distinguishable from the voiced distributions.

To investigate whether different unvoiced F0 generation approaches have a significant effect on the synthesised speech, two more listening tests were conducted in which the interpolation-based and random sampling F0 generation methods were compared. In both cases, the resultant HMM-GTD systems were compared to the default 1-Best F0 HMM-GTD system and the results are shown in figure 4.

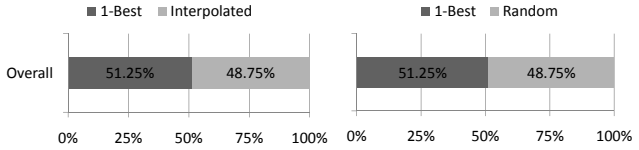


**Fig. 4**. Comparison between different uv F0 generation approaches

From figure 4, it is clear that there is little difference between the unvoiced F0 generation approaches ($p = 0.28$). This shows that the gain from the HMM-GTD model is obtained primarily from the model structure rather than the specific unvoiced F0 generation method used.

## 5. CONCLUSION

This paper has proposed a probabilistic modelling method for F0 values in unvoiced regions for HMM-based synthesis. The key idea is that F0 values are assumed to exist in unvoiced regions but they are drawn from a separate distribution. Experiments have been presented which show that a globally tied distribution (GTD) provides a good model for the unvoiced F0 regions. Furthermore, the resulting speech quality as measured by a subjective preference test is significantly better than that generated by the currently preferred multi-space distribution model (MSDHMM) . Also, it was shown that the approach is not sensitive to the method used for generating F0 training samples in unvoiced regions. Thus, overall the HMM-GTD appears to provide improved quality compared to an MSDHMM and since it is a regular HMM, it has the added benefit that existing algorithms such as adaptation can be applied without modification.

## 6. REFERENCES


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

[2] H. Kawahara, I. M. Katsuse, and A. D. Cheveigne, "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*, vol. 27, no. 3–4, pp. 187–207, 1999.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.

[4] S. Imai, "Cepstral analysis synthesis on the Mel frequency scale," in *Proc. ICASSP*, 1983.

[5] H. Kawahara, H. Katayose, A. D. Cheveigne, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity," in *Proc. EUROSPEECH*, 1999, pp. 2781–2784.

[6] D. Talkin, *Speech coding and synthesis*, chapter A robust algorithm for pitch tracking (RAPT), pp. 497–516, Elsevier, Ed., 1995.

[7] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[8] G. J. Freij and F. Fallside, "Lexical stress recognition using hidden Markov modeld," in *ICASSP*, 1988, pp. 135–138.

[9] U. Jensen, R. K. Moore, P. Dalsgaard, and B. Lindberg, "Modelling intonation contours at the phrase level using continous density hidden Markov models," *Computer Speech and Language*, vol. 8, pp. 247–260, 1994.

[10] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A pitch pattern modeling technique using dynamic features on the border of voiced and unvoiced segments," *Technical report of IEICE*, vol. 101, no. 325, pp. 53–58, 2001.

[11] T. Masuko, K. Tokuda, N. Miyazaki, and T. Kobayashi, "Pitch pattern generation using multi-space probability distribution hmm," *IEICE Trans.*, vol. J83-D-II, no. 7, pp. 1600–1609, 2000.

[12] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL principle for speech recognition," in *Proc. EUROSPEECH*, 1997, pp. 99–102.

[13] "HMM-based Speech Synthesis System (HTS)," http://hts.sp.nitech.ac.jp.