JOINT MODELLING OF VOICING LABEL AND CONTINUOUS F0 FOR HMM BASED SPEECH SYNTHESIS

K. Yu and S. Young

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK
Email: {ky219, sjy}@eng.cam.ac.uk

ABSTRACT

Fundamental frequency, or F0 is critical for high quality speech synthesis in HMM based speech synthesis. Traditionally, F0 values are considered to depend on a binary voicing decision such that they are continuous in voiced regions and undefined in unvoiced regions. Multi-space distribution HMM (MSDHMM) has been used for modelling the discontinuous F0. Recently, a continuous F0 modelling framework has been proposed and shown to be effective, where continuous F0 observations are assumed to always exist and voicing labels are explicitly modelled by an independent stream. In this paper, a refined continuous F0 modelling approach is proposed. Here, F0 values are assumed to be dependent on voicing labels and both are jointly modelled in a single stream. Due to the enforced dependency, the new method can effectively reduce the voicing classification error. Subjective listening tests also demonstrate that the new approach can yield significant improvements on the naturalness of the synthesised speech. A dynamic random unvoiced F0 generation method is also investigated. Experiments show that it has significant effect on the quality of synthesised speech.

Index Terms— HMM based speech synthesis, continuous F0 modelling, voicing classification

1. INTRODUCTION

In HMM based speech synthesis, the modelling of *Fundamental frequency (F0)* is difficult due to the discontinuity of F0 values. F0 is an inherent property of periodic signals and can represent the perceived *pitch* of human speech. During voiced speech, it is the periodic airflow modulation that serves as the excitation for the vocal tract. As there exists strong periodicity, F0 values can be effectively estimated from the waveform [1]. 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 [2] and hence the estimation is not reliable. Therefore, traditionally, F0 values during unvoiced region are assumed to be undefined, which leads to a discontinuous F0 observation stream. One widely used solution to directly model the discontinuous F0 observation is the *multi-space probability distribution HMM* (MSDHMM) [3]. Essentially, it uses a joint distribution of voicing label and discontinuous F0 observation as the state output distribution. Using the distribution of discontinuous F0 [4], HMM training can be performed efficiently and good performance can be achieved [5].

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).

Recently, an alternative model, continuous F0, has been investigated within the HMM based speech synthesis framework. Here, continuous F0 is assumed to exist in unvoiced regions and there have been a number of modelling approaches along this line. In [6], random F0 values are used in unvoiced regions and voicing labels are assumed to be hidden. A Gaussian mixture model (GMM) is employed, where unvoiced Gaussian components are globally tied so that the statistical difference between voiced and unvoiced regions can be modelled. In [7], voicing labels are assumed to be observable and modelled in an independent stream. As the voicing labels are explicitly modelled, global tying as defined in [6] is no longer a requirement for distinguishing voiced regions from unvoiced regions. Both approaches have shown significant improvement in the naturalness of synthesised speech compared to the traditional MSDHMM approach. This improvement is mainly due to the continuous F0 assumption [4] which leads to better modelling of the F0 trajectory. However, objective experiments have shown that the voicing classification performance of the proposed continuous F0 model are much worse than MSDHMM, which may potentially limit possible gains.

In this paper, a refined continuous F0 modelling approach is proposed. Here, continuous F0 is assumed to be dependent on observable voicing labels and they are modelled in the same stream. By enforcing the dependency, voicing classification performance can be improved. Experiments also show that this further improves the naturalness of the synthesised speech. One important issue in the continuous F0 framework is how to generate unvoiced F0 values. Though it has been shown that different unvoiced F0 generation approaches do not make much performance difference, a dynamic random generation approach is investigated in this paper. Experiments suggest that the unvoiced F0 generation approach can have significant effect on the synthesised speech quality.

The rest of the paper is arranged as follows. Section 2 compares different F0 modelling approaches. Section 3 describes the update formula of the proposed approach and some implementation issues. Objective and subjective experiments are presented in section 4, which is followed by conclusion.

2. COMPARISON OF F0 MODELLING APPROACHES FOR HMM BASED SPEECH SYNTHESIS

As described in section 1, there are two different fundamental assumptions regarding the F0 observations:

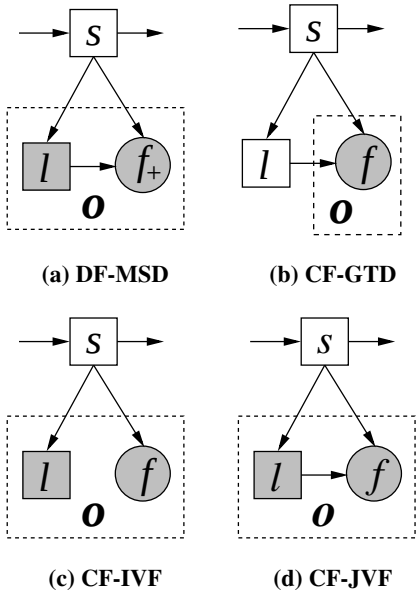
- *Discontinuous F0* assumes that the F0 observation is a real value in voiced regions but is a discrete symbol in unvoiced regions, referred to as f_+ in this paper:

$$f_+ \in \{\texttt{NULL}\} \cup (-\infty, \infty) \tag{1}$$

where NULL is the discrete symbol representing the observed F0 value in unvoiced regions. It is worth noting that NULL is not a voicing label, it is an *F0 observation value*.

- *Continuous F0* assumes real F0 value for all regions, referred to as $f \in (-\infty, \infty)$. Then the unvoiced F0 values have to be generated. They can be the 1-Best candidates from an F0 extractor, random samples or interpolated values between neighboring voiced regions [7].

Another important issue in F0 modelling is the modelling of the voicing label, referred to as $l \in \{\mathtt{U}, \mathtt{V}\}$, where U means unvoiced and V voiced. Different F0 modelling approaches can be compared according to the different assumptions and modelling of F0 values and voicing labels. Figure 1 shows the dynamic Bayesian networks [1] of the previously used approaches and the proposed method.



(a) DF-MSD **(b) CF-GTD**

(c) CF-IVF **(d) CF-JVF**

Fig. 1. DBN comparison between F0 modelling approaches.

A widely used F0 modelling approach is the *multi-space probability distribution HMM* (MSDHMM). It assumes discontinuous F0 (DF) and observable voicing labels, referred to as DF-MSD in figure 1(a). It can be shown that the state output distribution of MSDHMM can be written as [5, 4]:

$$p(\mathbf{o}|s) \;=\; p(l, f_+|s) \;=\; \begin{cases} P(\mathtt{U}|s) & l = \mathtt{U} \\ P(\mathtt{V}|s)\,\mathcal{N}(f_+|s, \mathtt{V}) & l = \mathtt{V} \end{cases} \quad (2)$$

where the observation $\mathbf{o} = [f \;\; l]$, $\mathcal{N}(\cdot)$ is a Gaussian distribution. Due to the discontinuity, it is not convenient to calculate dynamic features of F0 at the boundary between voiced and unvoiced regions. Though there are some exceptions [8, 4], the most widely used method is to model static and dynamic features in separate streams [9]. This common implementation limits the power of HMMs to model the F0 trajectory.

[1] A DBN is a graph that shows the statistical dependencies of random variables. In a DBN, a circle represents a continuous variable, a square represents a discrete variable, unshaded variables are hidden, and shaded variables are observed. The lack of an arrow from A to B indicates that B is conditionally independent of A. Note that for convenience the notation of continuous random variables is also used here for the discontinuous f_+.

Continuous F0 modelling is proposed to improve F0 trajectory modelling. By generating real F0 values for unvoiced regions and assuming hidden voicing labels, the Continuous F0 model with Globally Tied Distribution [6], *CF-GTD* in figure 1(b), is obtained. The state output distribution can be expressed as

$$p(\mathbf{o}|s) \;=\; p(f|s) \;=\; P(\mathtt{V}|s)\mathcal{N}(f|s, \mathtt{V}) + P(\mathtt{U}|s)\mathcal{N}(f|\mathtt{U}) \quad (3)$$

With continuous F0 values, it is easy to model static and dynamic features in a single stream. Experiments have shown that it can greatly reduce the F0 trajectory modelling error, and consequently improve the naturalness of the synthesised speech [6, 7]. However, due to hidden voicing labels, voicing classification only relies on the statistical difference between the globally tied unvoiced component and the state specific voiced component. This leads to significantly degraded voicing classification accuracy compared to MSDHMM.

To improve the voicing classification performance, voicing labels are assumed to be observable [7]. Here, an independent data stream is introduced to explicitly model voicing labels, referred to as Continuous F0 modelling with Independent Voicing label and F0 value, *CF-IVF* in 1(c). The corresponding state output distribution can be written as

$$p(\mathbf{o}|s) \;=\; p(l, f|s) \;=\; P(l|s)^{\gamma_l}\, p(f|s)^{\gamma_f} \quad (4)$$

where $p(f|s)$ and $P(l|s)$ are the distributions for the continuous F0 and voicing label streams respectively. $p(f|s)$ can be a single Gaussian or any other distribution such as a GMM. Dynamic features are only calculated for the continuous F0 values, not for voicing labels. γ_f and γ_l are stream weights. In practice, γ_f is set to be 1 and γ_l is set to be almost 0. This means, during HMM training, voicing labels do not contribute to the forward-backward state alignment stage but the model parameters are updated once the state alignment has been determined. In CF-IVF, the two streams share the same state clustering structure. During synthesis, state voicing status is only determined by the voicing label stream.

Though using observable voicing labels can improve voicing classification performance, it is still weak compared to MSDHMM due to the weak correlation between the two streams. In this paper, a refined approach is proposed, where only one stream is used to simultaneously model both observable voicing labels and continuous F0 values, referred to as Continuous F0 modelling with Joint Voicing label and F0 value, *CF-JVF* in 1(d). The state output distribution is

$$p(\mathbf{o}|s) \;=\; p(l, f|s) \;=\; P(l|s)\, p(f|s, l) \quad (5)$$

Compared to CF-IVF, CF-JVF introduces correlation between voicing labels and continuous F0 values and allows voicing labels to affect the forward-backward state alignment process. This will naturally strengthen the voicing label modelling. The DBN of CF-JVF is the same as MSDHMM. However, the observation definition is different. In MSDHMM, each observation dimension is a discontinuous variable as defined in equation (1). In contrast, CF-JVF uses different data types for different dimensions. Each dimension is either discrete or continuous, but not mixed. Only the continuous F0 dimensions require calculation of dynamic features.

3. JOINT MODELLING OF VOICING LABEL AND CONTINUOUS F0

The previous section introduces the motivation for jointly modelling voicing labels and continuous F0. This section will discuss the parameter update formula and implementation issues.

With equation (5) as the state output distribution, the auxiliary function for the parameters of voiced regions can be written as (irrelevant constants are ignored):

$$\mathcal{Q}(\mathcal{M}_{\mathtt{V}}) = \sum_s \sum_t \gamma_s(t)\delta(l_t, \mathtt{V})\Big(\log p(\boldsymbol{f}_t|s, \mathtt{V}) + \log P(\mathtt{V}|s)\Big) \quad (6)$$

where $\mathcal{M}_{\mathtt{V}}$ denotes the set of model parameters for the voiced part, $\boldsymbol{f} = [f\ \Delta f\ \Delta^2 f]$ denotes the continuous F0 vector consisting of static and dynamic features, l denotes voicing label, s denotes state and $\gamma_s(t)$ is the posterior probability of state s being at time t, $\delta(l_t, \mathtt{V})$ is a discrete delta function, whose value is 1 if $l_t = \mathtt{V}$ and 0, otherwise. From equation (6), the update formula for the parameters of $p(\boldsymbol{f}_t|s, \mathtt{V})$ is the same as the standard ML update formula except for using $\gamma_s(t)\delta(l_t, \mathtt{V})$ instead of $\gamma_s(t)$. In this paper, a single Gaussian is used as $p(\boldsymbol{f}_t|s, l)$, which leads to the following update formula

$$\boldsymbol{\mu}_{s,\mathtt{V}} = \frac{\sum_t \gamma_s(t)\delta(l_t, \mathtt{V})\boldsymbol{f}_t}{\sum_t \gamma_s(t)\delta(l_t, \mathtt{V})} \quad (7)$$

$$\boldsymbol{\Sigma}_{s,\mathtt{V}} = \frac{\sum_t \gamma_s(t)\delta(l_t, \mathtt{V})(\boldsymbol{f}_t - \boldsymbol{\mu}_{s,\mathtt{V}})}{\sum_t \gamma_s(t)\delta(l_t, \mathtt{V})} \quad (8)$$

The update formula for $p(\boldsymbol{f}_t|s, \mathtt{U})$ is similar. In this paper, all unvoiced distributions are tied as it is believed that the statistical properties of unvoiced F0 values should not be dependent on individual states. The probability of the voicing label $l \in \{\mathtt{U}, \mathtt{V}\}$ can be derived as

$$P(l|s) = \frac{\sum_t \gamma_s(t)\delta(l_t, l)}{\sum_t \gamma_s(t)} \quad (9)$$

Although the observation of CF-JVF consists of voicing label and continuous F0 value, during decision tree based state clustering, only the continuous F0 Gaussian is considered for convenience. With this approximation, the clustering process remains unchanged. During the synthesis stage, each state of the HMMs is classified as voiced or unvoiced state by comparing $P(l|s)$ to a predefined threshold (0.5 in this paper). In addition to CF-JVF, in this paper, a new unvoiced F0 generation method is also investigated. Here, samples from a pre-defined Gaussian distribution with large variance are generated as unvoiced F0 values. However, instead of one-off generation, those unvoiced F0 values are re-generated after each parameter estimation iteration. By introducing these *dynamic* random values for unvoiced regions, it is expected that the randomness of unvoiced F0 is better represented.

4. EXPERIMENTS

The performance of CF-JVF has been evaluated on two CMU ARCTIC speech synthesis data sets[10]. A U.S. female English speaker, `slt`, and a Canadian male speaker, `jmk`, were used. Each data set contains 1132 phonetically balanced sentences totalling about 0.95 hours of speech per speaker. To obtain objective performance measures, 1000 sentences from each data set were randomly selected for the training set, and the remainder were used to form a test set.

All systems were built using a modified version of the HTS toolkit [11]. Mixed excitation using STRAIGHT was employed [12]. The speech features used were 24 Mel-Cepstral spectral coefficients, the logarithm of F0, and aperiodic components in five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6 and 6 to 8 KHz). Spectral, F0 and aperiodic component features were modelled in separate streams during context-dependent HMM training. MDL-based state clustering [13] was performed for each stream to group the parameters of the context-dependent HMMs at state level. The MDL factor for MSDHMM is tuned so that it has a similar number of parameters as the continuous F0 modelling techniques. The same MDL factor is used for comparing CF-IVF and CF-JVF.

4.1. Objective comparison

To quantitatively compare discontinuous and continuous F0 modelling, the *root mean square error* (RMSE) of F0 observations and the *voicing classification error* (VCE) were calculated. The definition of RMSE and VCE can be found in [7].

Data Set	HMM	Female		Male	
		RMSE	VCE (%)	RMSE	VCE (%)
train	MSD	16.39	4.71	12.32	5.16
	CF-GTD	11.98	17.74	8.52	18.84
	CF-IVF	11.33	7.01	9.18	8.09
	CF-JVF	10.56	6.49	8.09	6.81
test	MSD	16.65	5.85	13.37	7.17
	CF-GTD	14.67	18.36	11.12	19.49
	CF-IVF	12.58	7.29	11.90	8.43
	CF-JVF	12.87	7.12	11.13	8.13

Table 1. Objective comparisons between F0 modelling approaches

From table 1, it can be observed that all continuous F0 approaches obtain significantly better RMSE on both training and test dataset than MSDHMM. However, VCE performance becomes worse when continuous F0 assumption is used. CF-GTD has the worst performance due to weak modelling of voicing labels. By explicitly modelling observable voicing labels, CF-IVF obtains significant improvement. The proposed CF-JVF approach can achieve further improvement on VCE due to the strengthened correlation between voicing label and continuous F0 values whilst retaining similar or better RMSE performance.

4.2. Subjective comparison

To confirm the results from objective experiments, a number of pairwise preference listening tests were conducted. As the comparison between continuous F0 modelling and MSDHMM has been given in [7], in this paper, the proposed CF-JVF approach is only compared to the previously best model, CF-IVF.

For the test material 30 sentences from a tourist information enquiry application were used. Two wave files were synthesised for each sentence and each speaker from the systems to be compared. Five sentences were then randomly selected to make up a test set for each listener, leading to 10 wave file pairs (5 male, 5 female). To reduce the noise introduced by forced choices, the 10 wave file pairs were duplicated and the order of the two systems were swapped. The final 20 wave file pairs were then shuffled and provided to the listeners in random order. Each listener was asked to select the more natural utterance from each wave file pair. Amazon mechanical turk is used to recruit listeners. Altogether 39 listeners, 25 native and 14 non-native, participated in the test. The result is shown in figure 2:

Statistical significance tests were performed for the result assuming a binomial distribution for each choice. The preference for CF-JVF was shown to be significant at 95% confidence level (p-

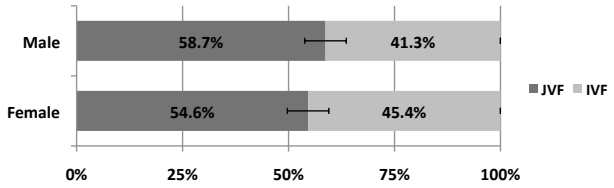


Fig. 2. Comparison between CF-IVF and CF-JVF on a forced choice preference test. Confidence interval of 95% is shown.

values: 0.03 for female and 0.0002 for male). This is consistent with the objective measures.

In the previous experiment, the 1-Best F0 candidate from the STRAIGHT F0 extractor was used as the unvoiced F0 value. Although different unvoiced F0 value generation approaches have in the past been shown to give similar performance [6, 7], all methods have been static. To reflect the randomness property of unvoiced F0, the dynamic random generation approach described above was investigated.

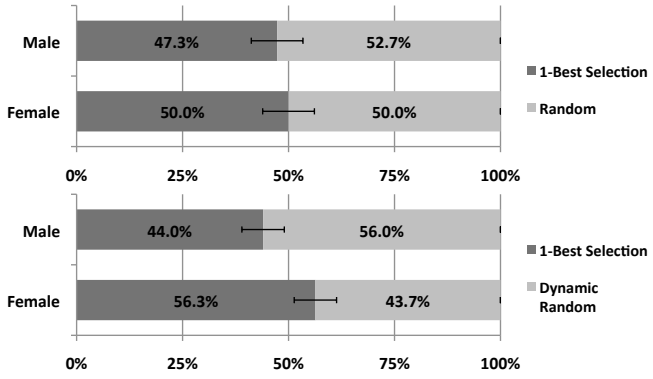


Fig. 3. Comparison between 1-best selection and random generation of unvoiced F0 values. Confidence interval of 95% is shown.

Figure 3 shows the comparison between the 1-Best selection approach and two random unvoiced F0 generation approaches. 26 listeners, 16 native and 10 non-native, participated in the first test, and 30 listeners, 18 native and 12 non-native participated in the second one. It can be observed that the speech quality of both speakers did not change much with static random F0 generation. Although the male speaker got degraded performance, it is not significant according to statistical significance test (p-value is 0.18). In contrast, dynamic random F0 generation showed enlarged difference. The male speaker was degraded while the female speaker was improved. Both were significant (p-value: 0.01 for female and 0.02 for male). This shows that different unvoiced F0 generation approaches do have an effect on the naturalness of the synthesised speech. However, the trend is different for the male and the female speaker. This is still to be investigated in future work.

5. CONCLUSION

This paper proposed a new continuous F0 modelling approach to strengthen the correlation between voicing labels and continuous F0 values. Continuous F0 values are dependent on the voicing labels and both are modelled in the same stream. Both objective measure and subjective listening tests showed that the proposed approach can achieve further improvement in the naturalness of synthesised speech, compared to the previous best continuous F0 modelling approach. A dynamic random unvoiced F0 generation method is also tested within the new framework. It is shown that unvoiced F0 generation can significantly affect the speech quality. However, the trend is not consistent. This will be a topic for future investigation.

6. REFERENCES

[1] 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.

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

[3] 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.

[4] K. Yu, B. Thomson, and S. Young, "From discontinuous to continuous f0 modelling in hmm-based speech synthesis," in *Proc. ISCA SSW7*, 2010.

[5] T. Yoshimura, *Simultaneous modelling of phonetic and prosodic parameters, and characteristic conversion for HMM based text-to-speech systems*, Ph.D. thesis, Nagoya Institute of Technology, 2002.

[6] K. Yu, T. Toda, M. Gasic, S. Keizer, F. Mairesse, B. Thomson, and S. Young, "Probablistic modelling of F0 in unvoiced regions in HMM based speech synthesis," in *Proc. ICASSP*, 2009.

[7] K. Yu and S. Young, "Continuous f0 modelling for hmm based statistical speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, 2010, To Appear.

[8] 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.

[9] 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.

[10] J. Kominek and A. Black, "CMU ARCTIC databases for speech synthesis," Tech. Rep. CMU-LTI-03-177, Language Technology Institute, School of Computer Science, Carnegie Mellon University, 2003.

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

[12] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *Proc. MAVEBA*, 2001.

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