From Discontinuous To Continuous F0 Modelling In HMM-based Speech Synthesis

Kai Yu, Blaise Thomson and Steve Young

Cambridge University Engineering Department, Trumpington Street, Cambridge, UK

{ky219,brmt2,sjy}@eng.cam.ac.uk

Abstract

The accurate modelling of *fundamental frequency, or F0,* in HMM-based speech synthesis is a critical factor in achieving high quality speech. However, it is also difficult because F0 values are normally considered to depend on a binary voicing decision such that they are continuous in voiced regions and undefined in unvoiced regions. A widely used solution is to use a multi-space probability distribution HMM (MSDHMM), which directly models discontinuous F0 observations. An alternative solution, continuous F0 modelling, has been recently proposed and shown to be more effective in achieving natural synthesised speech. Here, continuous F0 observations are assumed to always exist and hence they can be modelled by standard HMMs.

This paper describes a general mathematical framework for discontinuous F0 modelling, of which MSDHMM is a special case, and compares it to continuous F0 modelling. Various aspects associated with continuous F0 modelling, the use of a single F0 stream, globally tied distributions (GTD) and the assumption of a continuous unvoiced F0, are discussed in theory and examined in experiments. Both objective measures and subjective listening tests demonstrate that the introduction of continuous unvoiced F0 is vital for achieving speech quality improvement.

Index Terms: F0 modelling, MSDHMM, globally tied distribution, HMM based speech synthesis.

1. Introduction

HMM-based statistical speech synthesis has recently attracted much interest due to its compact and flexible representation of voice characteristics [1]. Based on the source-filter model assumption, phonetic and prosodic information are assumed to be conveyed primarily by the spectral envelope, fundamental frequency (also referred to as *F0*) and the duration of individual phones. A unified HMM framework may then be used to simultaneously model these features, where the spectrum and F0 are typically modelled in separate streams due to their different characteristics and time scales. During the synthesis stage, given a phone context sequence generated from text analysis, the corresponding sequence of HMMs are concatenated and spectral parameters and F0 are generated. These speech parameters are then converted to a waveform using synthesis filters.

The modelling of *fundamental frequency, or F0,* in HMM-based speech synthesis is a critical factor in delivering speech which is both natural and accurately conveys all of the many nuances of the message. However, F0 modelling is difficult due to the differing nature of F0 observations within voiced and unvoiced speech regions. In voiced speech regions, F0 values can be effectively estimated over a relatively short-time period. These F0 observations are continuous and normally range from 60Hz to 300Hz for human speech [2]. In unvoiced speech regions, the long term spectrum of turbulent airflow tends to be a weak function of frequency [3], which means that the identification of a single reliable F0 value is not possible. However, in most F0 modelling approaches, F0 is assumed to be observable for all time instances[1]. Hence, any practical F0 modelling approach must be capable of dealing with two issues:

- Classifying each speech frame as voiced or unvoiced;
- Modelling F0 observations, especially those in unvoiced speech regions.

A widely accepted assumption for F0 values in unvoiced speech regions is that they are *undefined* and must be denoted by a discrete unvoiced symbol. Consequently, F0 is a time-varying variable whose domain is partly continuous and partly discrete. This is referred to as a *discontinuous* variable in this paper[2]. Due to the mixed data types of the variable domain, discontinuous F0 values are not readily modelled by standard HMMs.

One solution is to directly model the discontinuous F0 observation and the *multi-space probability distribution HMM* (MSDHMM) was proposed for this purpose [5]. In [6], this discontinuous F0 distribution is interpreted as a mixture of two distributions for continuous and discrete values respectively. There is no explicit analysis of the relationship between voicing labels and discontinuous F0 observations. This interpretation using "a mixture of two distributions" can lead to the misunderstanding that the MSDHMM is a Gaussian mixture model (GMM). In this paper, a formal general mathematical framework is provided for discontinuous F0 HMM (DF-HMM) and the treatment of voicing labels is discussed explicitly. MSDHMM is shown to be a special case of DF-HMM. The state output distribution of MSDHMM is a joint distribution of observable voicing label and discontinuous F0 observation. The conditional probability of discontinuous F0 is then defined as a discrete probability within unvoiced regions, and a continuous density within voiced regions. Within the general DF-HMM framework, extensions of traditional MSDHMM are also discussed.

With a multi-space state-output distribution for discontinuous F0, HMM training can be efficiently performed and good

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). Thank Mark Gales for helpful discussion.

[1] Unobservable unvoiced F0 has also been investigated in [4].This is out of the scope of both discontinuous and continuous F0 frameworks, hence not discussed in this paper.

[2] Note that the "discontinuous" F0 in this paper does not just mean the lack of smoothness when viewed as a function of time. Real-value function can also be discontinuous in that sense. In this paper, the domain with mixed types of values is the essential property for being "discontinuous".

performance can be achieved [6]. However, there is still significant scope for improving F0 modelling accuracy. An alternative solution to discontinuous F0 modelling is to assume that continuous F0 observations also exist in unvoiced regions and use standard GMMs to model them. In [7], unvoiced F0 values are randomly generated and a two-component GMM with a globally tied distribution is then used to model the continuous F0 observations. Explicit voicing label modelling was added in [8] to form a complete *continuous F0* HMM (CF-HMM) framework. Compared to MSDHMM, CF-HMM has shown to be able to achieve better F0 trajectory modelling and significant improvement in the quality of synthesised speech [8, 7]. Besides generating continuous unvoiced F0 values, there are several techniques involved in CF-HMM which may contribute to the improved F0 modelling. They include the use of a single F0 stream to improve correlation modelling between static and dynamic F0 features, and the use of globally tied distribution (GTD) to absorb F0 extraction errors. As these techniques can also be applied to discontinuous F0 observations, it is then interesting to investigate them within the discontinuous F0 modelling framework. In this paper, theoretical and experimental comparisons between discontinuous and continuous F0 modelling are given in detail. Various techniques used in CF-HMM are discussed within the discontinuous F0 framework. Both objective and subjective tests showed that the introduction of continuous unvoiced F0 values is essential for achieving improved F0 modelling.

The rest of the paper is arranged as follows. Section 2 introduces a general framework of discontinuous F0 modelling, of which MSDHMM is a special case. The theoretical comparison between MSDHMM and CF-HMM, as well as using single-stream F0 and GTD for discontinuous F0 modelling, are discussed in section 3. Section 4 presents the results of both objective and subjective tests. Conclusions then follow. Finally, a derivation of the discontinuous F0 distribution is given in the appendix.

2. Discontinuous F0 modelling

As indicated in section 1, a common assumption is that F0 is observable for all time instances and it is a real value in voiced regions while undefined in unvoiced regions. Since F0 values are always considered as *observable*, a specific form of representation needs to be chosen for the *observations* in unvoiced regions. A natural representation is to use a discrete symbol. F0 is therefore a *discontinuous* variable, whose domain is partly discrete and partly continuous, which will be denoted 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* which must be introduced to satisfy the assumption that F0 is observable. Though it is normally determined by the voicing label output from a F0 extractor, it is different from a voicing label because it is a *singleton* only used for denoting an unvoiced F0 observation.

Having introduced f_+, it is necessary to define a proper distribution for it. Though the domain of f_+ is a mixture of a discrete symbol and real values, a distribution can still be defined using measure theory, as shown in the appendix. The distribution in this case is defined via the probability of events, A_{f_+}:

$$P(A_{f_+}) = \lambda^d \, \delta(f_+, \texttt{NULL}) + \lambda^c \int_{f_+ = f \in A_{f_+}} \mathcal{N}(f) \, df \tag{2}$$

where $f \in (-\infty, +\infty)$ denotes a real number, $\mathcal{N}(\cdot)$ is a Gaussian density of f, $\delta(\cdot, \cdot)$ is a discrete delta function defined as

$$\delta(a, b) \quad = \quad \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases}$$

$\lambda^d + \lambda^c = 1$ are prior probabilities of f_+ being discrete or continuous respectively and A_{f_+} is the event defined as:

$$A_{f_+} \quad = \quad \begin{cases} \texttt{NULL} & f_+ = \texttt{NULL} \\ (f, f + \Delta) & f_+ = f \in (-\infty, +\infty) \end{cases}$$

where Δ is a small interval. Equation (2) is a valid probability mass function. It is also possible to use a density-like form of equation (2) for the state output distribution in an HMM as follows

$$p(f_+) = \lambda^d \, \delta(f_+, \texttt{NULL}) + \lambda^c \, \mathcal{N}(f) \big(1 - \delta(f_+, \texttt{NULL})\big) \tag{3}$$

The use of the density form, equation (3), is equivalent to using the probability form, equation (2), during HMM training. Refer to the appendix for a more detailed explanation.

2.1. General form of discontinuous F0 HMM

As discussed above, the discrete symbol NULL is different from a voicing label which in this paper will be denoted explicitly as

$$l \in \{\texttt{U}, \texttt{V}\} \tag{4}$$

The issue here is that a typical F0 extractor which generates the observations used for model training makes both a voiced/unvoiced (V/U) decision represented as NULL and an estimate of real F0 values in voiced regions. However, there will be errors in the V/U decision and hence the true underlying voicing label must be regarded as being *hidden*. The output distribution of f_+ for state s should therefore be expressed as

$$\begin{aligned} p(f_+|s) \; &= \; P(\texttt{U}|s) p_{\texttt{u}}(f_+|s) + P(\texttt{V}|s) p_{\texttt{v}}(f_+|s) && (5) \\ &= \; (c^s_{\texttt{u}} \lambda^d_{\texttt{u}} + c^s_{\texttt{v}} \lambda^d_{\texttt{v}}) \delta(f_+, \texttt{NULL}) + \Big(c^s_{\texttt{u}} \lambda^c_{\texttt{u}} \mathcal{N}(f|s, \texttt{U}) + \\ & \quad c^s_{\texttt{v}} \lambda^c_{\texttt{v}} \mathcal{N}(f|s, \texttt{V}) \Big) \big(1 - \delta(f_+, \texttt{NULL})\big) && (6) \end{aligned}$$

where $P(\texttt{U}|s) = c^s_{\texttt{u}}$ and $P(\texttt{V}|s) = c^s_{\texttt{v}}$ are state dependent voicing probabilities subject to $c^s_{\texttt{u}} + c^s_{\texttt{v}} = 1$, $p_{\texttt{u}}(f_+|s)$ and $p_{\texttt{v}}(f_+|s)$ are conditional distributions of f_+, which take the form of equation (3) and lead to the form of equation (6).

By definition, $c^s_{\texttt{u}} \lambda^c_{\texttt{u}} \mathcal{N}(f|s, \texttt{U})$ is the likelihood contribution of the real F0 values detected within unvoiced regions. This term arises because the observed NULL symbol does not correspond exactly with the underlying voicing label l. It can be regarded as modelling erroneous voiced F0 values arising from a voicing classification error in the F0 extractor. Similarly, $c^s_{\texttt{v}} \lambda^d_{\texttt{v}}$ accounts for the error in misclassifying voiced speech as unvoiced. Therefore, equation (6) offers a complete framework for modelling both voicing classification and discontinuous F0 values. An HMM with equation (6) as its state output distribution is referred to as a *discontinuous F0 HMM* (DF-HMM). Once DF-HMMs are trained, they can be used for classifying the voicing condition of each state and generating voiced F0

parameters during synthesis. The state voicing classification can be naturally made by comparing $c^s_{\mathtt{v}}\lambda^c_{\mathtt{v}}$ to a predetermined threshold. Then, the voiced F0 parameters can be generated from $\mathcal{N}(f|s,\mathtt{V})$. One problem with this general form of DF-HMM is that voicing labels are hidden, hence the distinction between $\mathcal{N}(f|s,\mathtt{U})$ and $\mathcal{N}(f|s,\mathtt{V})$ relies solely on the difference in statistical properties between the erroneous F0 values and the correct F0 values, which could be hard to capture. This problem will be further discussed later.

2.2. Multi-space probability distribution HMM

The Multi-space probability distribution HMM (MSDHMM) is a special case of the DF-HMM in which voicing labels are assumed to be observable and the F0 extractor is assumed to be perfect. Therefore, the observation stream for the MSDHMM also includes the voicing label l and all terms modelling F0 extraction error will be zero

$$\lambda^c_{\mathtt{u}} = \lambda^d_{\mathtt{v}} = P(\mathtt{NULL}|\mathtt{V}) = 0 \quad (7)$$

$$\lambda^c_{\mathtt{v}} = \lambda^d_{\mathtt{u}} = P(\mathtt{NULL}|\mathtt{U}) = 1 \quad (8)$$

Equation (6) then becomes[3]

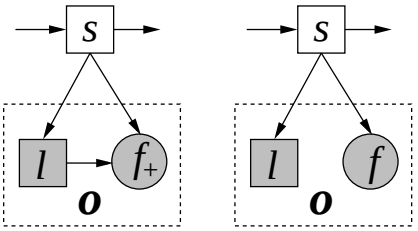
$$p(l,f_+|s) = P(l)p(f_+|l,s) = \begin{cases} c^s_{\mathtt{u}} & l=\mathtt{U} \\ c^s_{\mathtt{v}}\mathcal{N}(f|s,\mathtt{V}) & l=\mathtt{V} \end{cases} \quad (9)$$

where $c^s_{\mathtt{u}} + c^s_{\mathtt{v}} = 1$ are the prior voicing probabilities. In [6], equation (9) is interpreted as using different forms of distributions for discrete and continuous space respectively, which results in the name *multi-space* distribution. Though a GMM-like form is used in [6], it is worth noting that the state output distribution of the MSDHMM is not a mixture of expert model. From equation (9), it is clear that it is a joint distribution of voicing label and discontinuous F0 values, where due to the assumption of perfect F0 extraction, there will not be any cross-space terms. This approximation is convenient for both HMM training and voicing classification during synthesis. Hence, it has been widely used.

3. Comparison to continuous F0 modelling

Although the MSDHMM has achieved good performance, the use of discontinuous F0 has a number of limitations. Due to the discontinuity at the boundary between voiced and unvoiced regions, dynamic features can not be easily calculated and hence separate streams are normally used to model static and dynamic features [9]. This results in redundant voicing probability parameters which may not only limit the number of clustered states, but also weaken the correlation modelling between static and dynamic features. The latter would then limit the model's ability to accurately capture F0 trajectories. In addition, since all continuous F0 values are modelled by a single continuous density, parameter estimation is sensitive to voicing classification and F0 estimation errors. Furthermore, due to the nature of the discontinuous F0 assumption, one observation can only be either voiced or unvoiced, but not both at the same time. Consequently, during the forward-backward calculation 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 possibility of the unvoiced component to learn from voiced data and vice versa. Also, it forces the voiced component to be updated using all voiced observations making the system sensitive to F0 extraction errors.

To address these limitations, an alternative solution in the form of the continuous F0 HMM (CF-HMM), has been proposed [7, 8], where continuous F0 observations are assumed to also exist in unvoiced regions. In [7], voicing labels are assumed to be hidden, while in [8], observable voicing labels are used as in the MSDHMM.



(a) MSDHMM **(b) CF-HMM**

Figure 1: Dynamic Bayesian network comparison between MSDHMM and CF-HMM.

The dynamic Bayesian network comparison between MSDHMM and CF-HMM is shown in figure 1. In both cases, the observation includes the voicing label and the F0 values. In the MSDHMM, the voicing label and discontinuous F0 values are dependent, while in the CF-HMM, they are assumed to be independent since a separate stream is used to model the voicing classification [8]. The state output distribution of the CF-HMM is defined as

$$p(l,f|s) = p(f|s)^{\gamma_f} P(l|s)^{\gamma_l} \quad (10)$$

where $p(f|s)$ and $P(l|s)$ are the distributions for the continuous F0 and voicing label streams respectively, and γ_f and γ_l are stream weights. In [8], γ_f is set to be 1 and γ_l is set to be a very small positive value ϵ. During synthesis stage, the voicing condition is determined using the voicing label stream. Since there is no discontinuity in the continuous F0, it is possible to use any continuous density for $p(f|s)$. A two-component GMM, one state dependent and the other globally tied was used in [8].

Although the CF-HMM has been shown to yield significant improvement in speech quality compared to the MSDHMM [7, 8], it is not clear which aspects of the CF-HMM contribute most to the improvements. It is therefore useful to investigate the individual techniques used in the CF-HMM in more detail. The specific points of difference between the MSDHMM and the CF-HMM are:

1. *A single F0 stream* is used for both static and dynamic F0 features to provide a consistent voicing label probability and strong temporal correlation modelling.
2. *A globally tied distribution (GTD)* is used to yield robust unvoiced F0 estimation.
3. *The continuous F0 assumption* avoids the problem of modelling a discontinuity at V/UV boundaries. This allows a single F0 stream to be used and it also avoids the hard assignment of state posterior during HMM training.

[3] Strictly speaking, $\delta(\cdot,\cdot)$ should appear in equation (9) to denote that, under the MSDHMM assumption, it is not possible to observe $(\mathtt{U}, f)$ or $(\mathtt{V}, \mathtt{NULL})$. This is omitted for clarity.

It is interesting to note that only the continuous F0 assumption is an inherent property of CF-HMM. A single F0 stream can also be obtained for MSDHMM by constructing dynamic F0 features at unvoiced/voiced boundaries. For example, in [10], the boundary dynamic F0 features are calculated from the nearest voiced F0 observations across the unvoiced segment. It is then possible to use a single stream for both static and dynamic F0 features as they have the same voicing boundary. GTD is also not intrinsic to the CF-HMM. From the general DF-HMM, equation (6), GTD can be easily introduced. Assuming the F0 extraction error is independent of states and combining the prior weights together, equation (6) becomes

$$p(f_+|s) = c_1^s\delta(f_+,\texttt{NULL}) + \Big(c_2^s\mathcal{N}(f|\texttt{U}) + c_3^s\mathcal{N}(f|s,\texttt{V})\Big)(1-\delta(f_+,\texttt{NULL})) \quad (11)$$

and $c_1^s = c_{\texttt{u}}^s\lambda_{\texttt{u}}^d + c_{\texttt{v}}^s\lambda_{\texttt{v}}^d$, $c_2^s = c_{\texttt{u}}^s\lambda_{\texttt{u}}^c$, $c_3^s = c_{\texttt{v}}^s\lambda_{\texttt{v}}^c$, $c_1^s + c_2^s + c_3^s = 1$.

Given that a single F0 stream and GTD can both be implemented within the DF-HMM framework, the MSDHMM can be extended to include these and thereby allow a direct comparison with the CF-HMM. To use a single F0 stream, SPLINE interpolation is first performed for all unvoiced segments and dynamic real-valued F0 features are then constructed at the unvoiced/voiced boundaries. Consequently, a single F0 stream can be used to model the discontinuous F0 vectors, which are partly discrete NULL symbols and partly three-dimensional real-valued vectors (here only first and second derivatives are used). Furthermore, the GTD technique can be applied to the single stream MSDHMM. A globally tied Gaussian component is used as $\mathcal{N}(f|\texttt{U})$ in equation (11) and c_1^s, c_2^s, c_3^s are updated independently given the sum-to-one constraint. The GTD component is initialized using all voiced F0 values and is never updated during HMM training[4]. During synthesis, c_1^s is compared to a pre-determined threshold (0.5 in this paper) to determine the voicing classification for each synthesis frame.

4. Experiment

The comparison between the extended MSDHMM and the CF-HMM has been made using two CMU ARCTIC speech synthesis data sets [11]: the U.S. female English speaker slt, and the Canadian male speaker, jmk. Each data set contains recordings of the same 1132 phonetically balanced sentences totalling about 0.95 hours of speech per speaker. All systems were built using a modified version of the HTS HMM speech synthesis toolkit version 2.0.1 [12]. The feature set includes 24 spectral coefficients, log F0 and 5 aperiodic component features. Minimum description length (MDL) based state clustering, single Gaussian duration modelling, mixed excitation and global variance for synthesis were used. More details of this experimental set-up can be found in [8]. In this paper, the CF-HMM used explicit voicing condition modelling, where an extra data stream was used for voicing labels. To make a fair comparison, during state-clustering, the MDL scaling factors were tuned so that all systems have similar numbers of clustered F0 states.

[4] Additional experiments showed that updating the GTD component will lead to worse performance. This is because the parameters of the GTD will be heavily affected by the dominant voiced F0 data during training. Consequently, the updated GTD component will have a small variance although globally tied. This GTD will then fail to model outliers of voiced F0 and will adversely affect the training and state clustering process.

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 for each of the extended MSDHMM systems and the CF-HMM system. To obtain these objective measures, 1000 sentences from each data set were randomly selected for the training set, and the remainder were used to form a test set. To reduce the effect of the duration model when comparing the generated F0 trajectories, state level durations were first obtained by force-aligning the known natural speech from the test set. Then, given the natural speech durations, a voicing classification was performed for each state, followed by F0 value generation within the voiced regions. By this mechanism, natural speech and synthesised speech were aligned and could be compared frame by frame to give a root mean square error defined as

$$\texttt{RMSE} = \sqrt{\frac{\sum_{t\in\mathcal{V}}\left(f(t)-f_{\texttt{r}}(t)\right)^2}{\#\mathcal{V}}} \quad (12)$$

where $f_{\texttt{r}}(t)$ is the extracted F0 observation of the natural speech at time t, $f(t)$ is the synthesized F0 value at time t, $\mathcal{V} = \{t : l(t) = l_{\texttt{r}}(t) = \texttt{V}\}$ denotes the time indices when both natural speech and synthesized speech are voiced, $\#\mathcal{V}$ is the total number of voiced frames in the set. The voicing classification error is defined as the percentage of mismatched voicing labels

$$\texttt{VCE} = 100\frac{\sum_{t=1,T}\Big(1-\delta\big(l(t),l_{\texttt{r}}(t)\big)\Big)}{T} \quad (13)$$

where $\delta(l,l_r)$ is defined in equation (3), and T is the total number of frames. From table 1, it can be seen that compared to the standard MSDHMM, the single stream MSDHMM (MSD+1str) can slightly reduce the average F0 synthesis errors (RMSE) in both training and test sets presumably due to better temporal correlation modelling. However, it is still less accurate than the CF-HMM. The use of the GTD technique in the MSDHMM led to the worst RMSE performance. This shows that the GTD component cannot accurately capture F0 extraction errors. Instead, it will spoil the estimation of the other voiced Gaussian component because it can absorb mass from real-valued F0 observations in voiced regions. In contrast to the MSDHMM, the CF-HMM has randomly generated unvoiced F0 values which provide a strong statistical constraint (especially in the dynamic features) which prevents the GTD component from subsuming the correctly estimated voiced F0 observations. Hence, although the GTD can absorb F0 outliers

Data Set	HMM	Female		Male	
		RMSE	VCE (%)	RMSE	VCE (%)
train	MSD	16.14	4.48	12.00	4.90
	+ 1str	15.94	5.76	11.53	6.68
	+ GTD	21.19	5.44	19.09	6.51
	CF	11.33	7.01	9.18	8.09
test	MSD	16.76	5.85	13.34	6.90
	+ 1str	15.77	6.85	12.79	8.26
	+ GTD	23.44	7.06	20.25	8.10
	CF	12.58	7.29	11.90	8.43

Table 1: Objective comparison between MSDHMM extensions and CF-HMM

and yield robust F0 estimation in the CF-HMM, it cannot do the same for the MSDHMM[8]. It is worth noting that from the definition of RMSE in this paper, equation (12), only the F0 values well inside voiced regions are considered. This implies that GTD with the continuous F0 assumption does not only apply to boundary observations, it also effectively applies to normal voiced speech regions. In terms of voicing classification error, all discontinuous F0 HMM approaches obtained better results than the CF-HMM. This is expected since the CF-HMM assumes independence between voicing label and F0 observations, hence the voicing label modelling is weaker. In particular, MSDHMM yielded the best VCE performance because it not only assumes observable voicing labels, but also assumes dependency between F0 observations and voicing labels.

4.2. Subjective listening tests

The previous objective measures show useful comparison information. This section will give the results of subjective listening tests to properly measure the effective performance of the different synthesis models.

The subjective evaluation consisted of a pair-wise preference test. For the test material, 30 sentences from a tourist information enquiry application were used. These sentences have quite different text patterns compared to the CMU ARCTIC text corpus and they therefore provide a useful test of the generalization capability of the systems. Two wave files were synthesised for each sentence and each speaker, one from the CF-HMM system and the other from the MSDHMM system. 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 waves were then shuffled and presented to the listeners in random order. Each listener was asked to select the more natural utterance from each wave file pair.

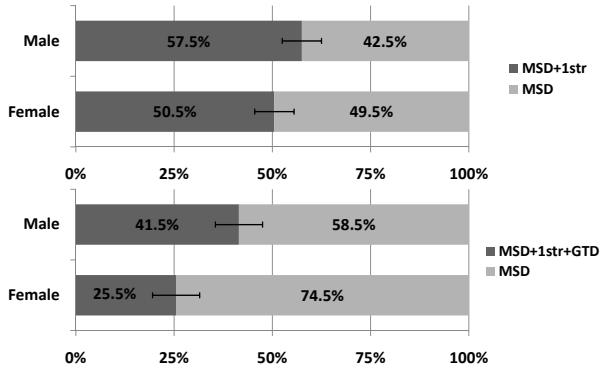


Figure 2: MSDHMM v.s. extended MSDHMM. Confidence Interval of 95% is shown.

Figure 2 shows the comparison between the two extended MSDHMM systems and the traditional MSDHMM. 8 native and 12 non-native listeners conducted the tests. As can be seen, the results are largely consistent with the objective measures. Using a single F0 stream improved the temporal correlation modelling and resulted in better synthesised speech. It can be observed that the effect on the male speaker is much stronger than the female speaker. Statistical significance tests show that the improvement on the quality of the male speech is significant at a 95% confidence level. For the female voice, there is almost no difference when using a single F0 stream. In contrast, adding GTD to the single F0 stream MSD system significantly degraded the quality of synthesised speech for both voices. This shows that GTD alone is not directly useful within the MSDHMM framework.

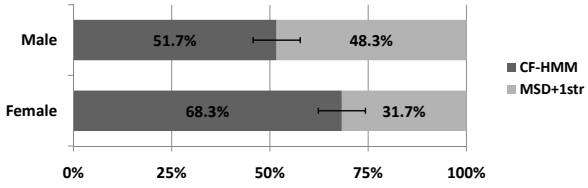


Figure 3: CF-HMM v.s. MSDHMM with single F0 stream. Confidence Interval of 95% is shown.

Figure 3 shows the comparison between CF-HMM and MSDHMM with a single F0 stream, which outperformed the traditional MSDHMM. 8 native and 10 non-native listeners participated in the test. As can be seen, the CF-HMM outperformed the MSDHMM with a single F0 stream. The improvement for the female voice is significant while insignificant for the male voice. This is expected since the single F0 stream MSDHMM achieved a significant improvement for the male voice compared to the standard MSDHMM. The only difference between the two systems in figure 3 is that the CF-HMM uses GTD with continuous F0 values, whilst the MSDHMM uses discontinuous F0 values. This shows that the continuous F0 assumption is an important factor in enabling the CF-HMM to achieve performance improvements.

5. Conclusion

F0 modelling in HMM-based speech synthesis is important. Due to the nature of undefined F0 in unvoiced regions, F0 is normally considered as a discontinuous variable, which is partly continuous and partly discrete. This paper describes a mathematical framework for modelling discontinuous F0 in HMM-based statistical speech synthesis (DF-HMM). The multi-space probability distribution HMM (MSDHMM) is then shown to be a special case of the DF-HMM.

The DF-HMM is then compared to the continuous F0 HMM (CF-HMM) framework, where real-valued F0 observations are assumed to exist in unvoiced regions. Since in addition to the inherent continuous F0 assumption, a single F0 stream and a globally tied distribution (GTD) are also used in CF-HMM, the MSDHMM is extended with both of these to allow it to be compared directly with the CF-HMM. Both objective measures and subjective listening tests showed that using a single F0 stream can improve F0 modelling, while the continuous F0 assumption is critical to the ability of the GTD technique to work effectively and achieve an overall improvement compared to the MSDHMM.

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] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Prentice Hall PTR, 2001.

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

[4] K. N. Ross and M. Ostendorf, "A dynamical system model for generating fundamental frequency for speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 295–309, 1999.

[5] 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.

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

[7] 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.

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

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

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

[13] A. Papoulis, *Probability, random rariables, and stochastic processes*. McGraw-Hill, 1984.

[14] W. Rudin, *Real and complex analysis, 3rd ed.* New York, NY, USA: McGraw-Hill, Inc., 1987.

7. Appendix

The definition of $p(f_+)$ follows the standard approach for distributions of mixed type (discrete and continuous). [13] provides discussions on the use of mixed distributions. A short discussion is included below for completeness. All terms used in this appendix are discussed in [14].

To define the probability distribution via measure theory, one must first define the collection of measurable events, called the σ-algebra. In the case discussed here the σ-algebra is the smallest σ-algebra containing the open intervals and also containing the singleton NULL (This exists by Theorem 1.10 of [14]). The probability measure, P, is defined in terms of the events, A. For values $a, b \in \mathbb{R}, a < b$, the probability function is defined as:

$$P(A) = \begin{cases} \lambda^d & A = \{\texttt{NULL}\} \\ \lambda^c \int_{f \in (a,b)} \mathcal{N}(f)\, df & A = (a,b) \end{cases},$$

where $\lambda^d + \lambda^c = 1$. Note that the probability function has only been defined in terms of open intervals and the $\{\texttt{NULL}\}$ singleton. This is sufficient because the σ-algebra used is the smallest σ-algebra containing these sets.

Despite the use of a mixed distribution, a probability density function may still be defined by using Lebesque integration. The corresponding probability function is defined as a function of $f_+ \in \{\texttt{NULL}\} \cup (-\infty, \infty)$ by:

$$p(f_+) = \lambda^d\, \delta(f_+, \texttt{NULL}) + \lambda^c\, \mathcal{N}(f)(1 - \delta(f_+, \texttt{NULL})). \quad (14)$$

This form of density function can be used in likelihood calculations during HMM training as if it were a standard density function.

To formalize the use of this function, one requires a measure to integrate over. Let the measure μ be defined as follows (with $a, b \in \mathbb{R}, a < b$):

$$\mu(\{\texttt{NULL}\}) = 1, \quad (15)$$

$$\mu((a,b)) = (b - a). \quad (16)$$

Using Lebesgue integration [14] of the probability density p, equation (14), with respect to this measure gives that:

$$P(A) = \int_A p\, d\mu. \quad (17)$$

Substituting in for the event A, the above formula in terms of traditional integration becomes (with $a, b \in \mathbb{R}, a < b$):

$$P(\{\texttt{NULL}\}) = p(\texttt{NULL}) = \lambda^d, \quad (18)$$

$$P((a,b)) = \int_{f \in (a,b)} p(f) df, \quad (19)$$

$$= \lambda^c \int_{f \in (a,b)} N(f) df. \quad (20)$$