Chapter 9
Statistical Models for Dealing with Discontinuity of Fundamental Frequency

Kai Yu

Abstract The accurate modelling of *fundamental frequency*, or *F0*, in HMM-based speech synthesis is a critical factor for 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. Namely, estimated F0 value is a discontinuous function of time, whose domain is partly continuous and partly discrete. This chapter investigates two statistical frameworks to deal with the discontinuity issue of F0. *Discontinuous F0 modelling* strictly defines probability of a random variable with discontinuous domain and model it directly. A widely used approach within this framework is *multi-space probability distribution* (MSD). An alternative framework is *continuous F0 modelling*, where continuous F0 observations are assumed to always exist and voicing classification is modelled separately. Both theoretical and experimental comparisons of the two frameworks are given.

9.1 Statistical Parametric Speech Synthesis with Discontinuous Fundamental Frequency (F0)

Compared to traditional unit concatenation speech synthesis approaches, statistical parametric speech synthesis has recently attracted much interest due to its compact and flexible representation of voice characteristics. Hidden Markov model (HMM)-based synthesis (Yoshimura et al. 1999) is the most widely used approach of statistical parametric speech synthesis and is the focus of this chapter. 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. The spectral and F0 features can be extracted from a speech waveform (Kawahara et al. 1999a), and durations can be manually labelled or obtained through forced-alignment using pre-trained HMMs. A unified

K. Yu (✉)
Computer Science and Engineering, Shanghai Jiao Tong University, 800,
Dongchuan Road, Minhang District, Shanghai, P. R. China
e-mail: kai.yu@sjtu.edu.cn

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_9

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 (Tokuda et al. 2000). These speech parameters are then converted to a waveform using synthesis filters (Imai 1983).

The modelling of F0 is difficult due to the differing nature of F0 observations within voiced and unvoiced speech regions. F0 is an inherent property of periodic signals and in human speech it represents the perceived *pitch*. During *voiced* speech such as vowels and liquids, the modulated periodic airflow emitted from the glottis serves as the excitation for the vocal tract. Since there is strong periodicity, F0 values can be effectively estimated over a relatively short-time period (e.g. a speech frame of 25 ms) using (Kawahara et al. 1999b). These F0 observations are continuous and normally range from 60 to 300 Hz for human speech (Huang et al. 2001). However, in *unvoiced* speech such as consonants, energy 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 (Talkin 1995), which means that the identification of a single reliable F0 value in unvoiced regions is not possible. However, in most F0 modelling approaches, F0 is assumed to be observable for all time instances[1]. Consequently, any practical F0 modelling approach must be capable of dealing with two issues:

- **Voicing classification**: classify each speech frame as voiced or unvoiced;
- **F0 observation representation**: model F0 observations in both voiced and unvoiced speech regions.

Voicing classification is often performed during F0 extraction (Kawahara et al. 1999b), and hence, the voicing label of each frame is usually assumed to be observable. Since the nature of each F0 observation depends on the type of voicing condition, voicing labels are normally considered together with F0 observations rather than being separately modelled. 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. Note that the "discontinuous" F0 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. Here, the domain with mixed types of values is the essential property for being "discontinuous". Due to the mixed data types of the variable domain, discontinuous F0 observations are not readily modelled by standard HMMs. *Discontinuity* of F0 is, therefore, an essential problem to address in HMM based speech synthesis.

[1] Unobservable unvoiced F0 has also been investigated in (Ross and Ostendorf 1999).This is out of the scope of both discontinuous and continuous F0 frameworks, hence not discussed here.

One solution is to directly model discontinuous F0 observations. The *multi-space probability distribution HMM* (MSDHMM) was proposed for this purpose (Tokuda et al. 2002). In (Yoshimura 2002), 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 a misunderstanding that the MSDHMM is a Gaussian mixture model (GMM). In this chapter, a formal general mathematical framework is provided for *discontinuous F0* HMM (DF-HMM) (Yu et al. 2010) and the treatment of voicing labels is discussed explicitly. MSDHMM is shown to be a special case of DF-HMM. 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 performance can be achieved (Yoshimura 2002). 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. This is referred to as *continuous F0* HMM (CF-HMM) framework. A number of approaches with different independency assumption between voicing label and F0 observation have been proposed (Yu et al. 2009; Yu and Young 2011a, b).

The rest of this chapter will use consistent mathematical notations to describe the two F0 modelling frameworks in detail. The two frameworks are then compared in both theory and experiments.

9.2 Discontinuous F0 Modelling

As indicated in Sect. 9.1, a common assumption is that F0 is observable for all time instances and it has 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_+:

$$f_+ \in \{\mathtt{NULL}\} \cup (-\infty, \infty), \tag{9.1}$$

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 can be 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 probability 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_+}:

$$P(A_{f_+}) = \lambda^d\, \delta(f_+, \texttt{NULL}) + \lambda^c \int_{f_+ = f \in A_{f_+}} \mathcal{N}(f)\, df \tag{9.2}$$

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

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

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

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

where Δ is a small interval. Equation (9.2) is a valid probability mass function. It is also possible to use a density-like form of Eq. (9.2) for the state output distribution in an HMM as follows

$$p(f_+) = \lambda^d\, \delta(f_+, \texttt{NULL}) + \lambda^c\, \mathcal{N}(f)(1 - \delta(f_+, \texttt{NULL})) \tag{9.3}$$

The use of the density form, Eq. (9.3), is equivalent to using the probability form, Eq. (9.2), during HMM training. Refer to the appendix for a more detailed explanation.

9.2.1 General Form Of Discontinuous F0 HMM

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

$$l \in \{\texttt{U}, \texttt{V}\} \tag{9.4}$$

The issue here is that a typical F0 extractor usually outputs a *single* observation stream representing both voicing (V/U) decision and the estimate of real F0 values in voiced regions. Although the voicing decision of F0 extractors is reflected by the switching between NULL and real F0 values, it is not guaranteed to give the real voicing boundaries due to voicing classification errors. Hence, the voicing label is assumed to be *hidden* and the output distribution of f_+ for state s should, therefore, be expressed as

$$p(f_+|s) = P(\texttt{U}|s)p_{\text{u}}(f_+|s) + P(\texttt{V}|s)p_v(f_+|s) \tag{9.5}$$

$$= (c_{\mathrm{U}}^s \lambda_{\mathrm{U}}^d + c_{\mathrm{V}}^s \lambda_{\mathrm{V}}^d)\delta(f_+, \mathtt{NULL}) + \big(c_{\mathrm{U}}^s \lambda_{\mathrm{U}}^c \mathcal{N}(f|s, \mathrm{U}) \\ + c_{\mathrm{V}}^s \lambda_{\mathrm{V}}^c \mathcal{N}(f|s, \mathrm{V})\big)(1 - \delta(f_+, \mathtt{NULL})), \tag{9.6}$$

where $P(\mathrm{U}|s) = c_{\mathrm{U}}^s$ and $P(\mathrm{V}|s) = c_{\mathrm{V}}^s$ are state dependent voicing probabilities subject to $c_{\mathrm{U}}^s + c_{\mathrm{V}}^s = 1$, $p_{\mathrm{U}}(f_+|s)$ and $p_{\mathrm{V}}(f_+|s)$ are conditional distributions of f_+, which take the form of Eq. (9.3) and lead to the form of Eq. (9.6).

By definition, $c_{\mathrm{U}}^s \lambda_{\mathrm{U}}^c \mathcal{N}(f|s, \mathrm{U})$ 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 to the underlying voicing label l. It can be regarded as modelling erroneous voiced F0 values arising from a voicing classification error in an F0 extractor. Similarly, $c_{\mathrm{V}}^s \lambda_{\mathrm{V}}^d$ accounts for the error in misclassifying voiced speech as unvoiced. Therefore, Eq. (9.6) offers a complete framework for modelling both voicing classification and discontinuous F0 values. An HMM using Eq. (9.6) as its state output distribution is referred to as a *discontinuous F0 HMM* (DF-HMM) (Yu et al. 2010). 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 $c_{\mathrm{V}}^s \lambda_{\mathrm{V}}^c$ to a predetermined threshold. Then, the voiced F0 parameters can be generated from $\mathcal{N}(f|s, \mathrm{V})$. One problem with this general form of DF-HMM is that voicing labels are hidden, hence the distinction between $\mathcal{N}(f|s, \mathrm{U})$ and $\mathcal{N}(f|s, \mathrm{V})$ relies solely on the difference in statistical properties between the erroneous F0 values and the correct F0 values, which could be hard to capture.

9.2.2 Multi-Space Probability Distribution HMM

MSDHMM is a special case of 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

$$\lambda_{\mathrm{U}}^c = \lambda_{\mathrm{V}}^d = P(\mathtt{NULL}|\mathrm{V}) = 0 \tag{9.7}$$

$$\lambda_{\mathrm{V}}^c = \lambda_{\mathrm{U}}^d = P(\mathtt{NULL}|\mathrm{U}) = 1 \tag{9.8}$$

Eq. (9.6) then becomes[2]

$$p(\mathbf{o}|s) = p(l, f_+|s) = P(l)p(f_+|l, s) = \begin{cases} c_{\mathrm{U}}^s & l = \mathrm{U} \\ c_{\mathrm{V}}^s \mathcal{N}(f|s, \mathrm{V}) & l = \mathrm{V} \end{cases} \tag{9.9}$$

where $c_{\mathrm{U}}^s + c_{\mathrm{V}}^s = 1$ are the prior voicing probabilities. In (Yoshimura 2002), Eq. (9.9) is interpreted as using different forms of distributions for discrete and continuous

[2] Strictly speaking, $\delta(\cdot, \cdot)$ should appear in Eq. (9.9) to denote that, under the MSDHMM assumption, it is not possible to observe (U, f) or $(\mathrm{V}, \mathtt{NULL})$. This is omitted for clarity.

space respectively, which results in the name *multi-space* distribution. Though a GMM-like form is used in (Yoshimura 2002), it is worth noting that the state output distribution of the MSDHMM is not a mixture of expert model. From Eq. (9.9), 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. The parameter estimation formula and details of using MSDHMM during synthesis stage can be found in (Tokuda et al. 2002).

9.3 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 cannot be easily calculated and hence separate streams are normally used to model static and dynamic features (Masuko et al. 2000). 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 HMM 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, *continuous F0 modelling*, is proposed (Yu et al. 2009; Yu and Young 2011a, b). In this framework, continuous F0 observations are assumed to exist in both voiced and unvoiced speech regions and hence both F0 and the voicing labels can be modelled by regular HMMs, referred to as *continuous F0 HMM* (CF-HMM).

Figure 9.1 shows the relationship between discontinuous and continuous F0 modelling where Fig. 9.1a represents the discontinuous case. As it can be seen, continuous F0 assumes real F0 value for all regions, i.e.

$$f \in (-\infty, \infty). \tag{9.10}$$

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 neighbouring voiced regions (Yu and Young 2011a). Another important issue is the modelling of voicing label, referred to as $l \in \{\mathtt{U}, \mathtt{V}\}$, where $\mathtt{U}$ means unvoiced and $\mathtt{V}$ voiced. It

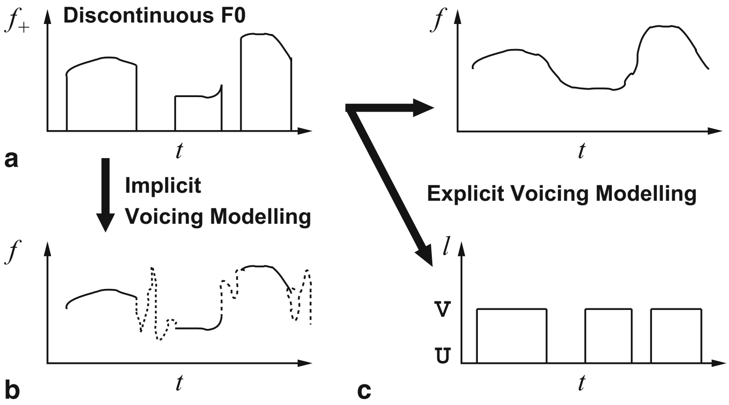


Fig. 9.1 Relationship between discontinuous F0 modelling (**a**) and continuous F0 modelling with implicitly determined voicing condition (**b**) and explicitly determined voicing condition (**c**)

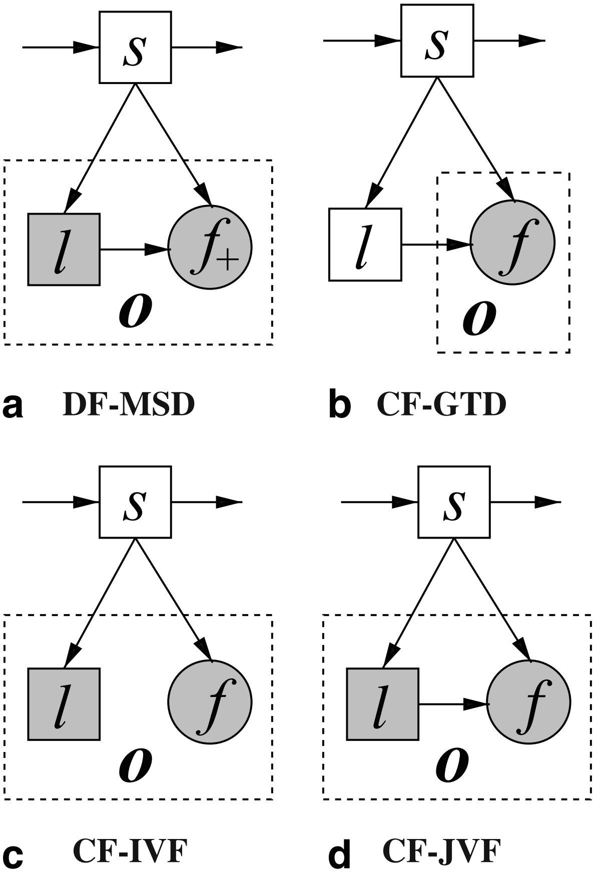
can be modelled as a hidden variable (implicitly) or an observable variable (explicitly). Different treatments of voicing labels lead to different CF-HMM approaches. Figure 9.2 shows the dynamic Bayesian networks [3] of MSDHMM and various continuous F0 approaches. The upcoming sections discuss various aspects of continuous F0 modelling in detail.

9.3.1 Determining F0 in Unvoiced Regions

If F0 is considered to exist in unvoiced regions, then there must, in practice, be some method of determining it. One approach is to make use of the pitch tracker used in F0 observation extraction, such as STRAIGHT (Kawahara et al. 1999a). In many pitch trackers, multiple F0 candidates are generated for each speech frame regardless of whether it is voiced or unvoiced. A post-processing step is then used to assign voicing labels. For voiced regions, the 1-best F0 candidates are reliable. They normally have strong temporal correlation with their neighbours and form a smooth trajectory. In contrast, for unvoiced regions, the 1-best F0 candidates do not have strong temporal correlation and tend to be random. The 1-best F0 candidates of unvoiced regions can, therefore, be used as F0 observations. This will be referred to as *1-best selection*.

[3] 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_+.

Fig. 9.2 DBN comparison between F0 modelling approaches



Note that unvoiced F0 observations near the boundaries of voiced regions may have temporal correlation which is useful when calculating dynamic features.

Other methods of determining F0 in unvoiced regions may also be used, such as sampling from a pre-defined distribution with large variance (Freij and Fallside 1988; Yu et al. 2009), using SPLINE interpolation (Lyche and Schumaker 1973) or choosing the F0 candidate which is closest to the interpolated F0 trajectory (Yu et al. 2009). Instead of one-off generation, dynamic random generation can also be used, where unvoiced F0 values are regenerated after each parameter estimation iteration (Yu and Young 2011b). It has been observed in various experiments, although synthesis quality may be occasionally affected, there is no consistent conclusion that one particular unvoiced F0 generation approach is best (Yu and Young 2011b; Yu et al. 2009). Hence, only 1-best selection is used in experiments of this chapter.

9.3.2 Different Forms of Continuous F0 Modelling

As indicated before, voicing label can be modelled either implicitly or explicitly. There can also be different assumptions on the dependency between F0 observations and voicing labels. These lead to different forms of continuous F0 modelling.

9.3.2.1 Continuous F0 Modelling with Globally Tied Distribution

Here, implicit voicing assumption is used, i.e. voicing label is not observable. By generating real F0 values for unvoiced regions and assuming hidden voicing labels, the continuous F0 modelling with globally tied distribution (Yu et al. 2009), *CF-GTD* in Fig. 9.2b, is obtained. If a frame is voiced then the extracted F0 value is used as the observation, otherwise some other method of computing F0 is used to derive the observation as discussed in the previous section [4].

As voicing labels are assumed to be hidden, a GMM (normally two-component) is used to model the continuous F0 observation f, with one component corresponding to voiced F0 and the other corresponding to unvoiced F0. Due to the uncorrelated nature of unvoiced F0 observations, the distribution of unvoiced F0 is assumed to be independent of the HMM states. The output distribution of an observation $\mathbf{o}$ at state s can then be written as

$$\begin{aligned} p(\mathbf{o}|s) = p(f|s) &= \sum_{l\in\{\mathrm{U},\mathrm{V}\}} P(l|s)p(f|l,s) \\ &= P(\mathrm{U}|s)\mathcal{N}(f;\mu_{\mathrm{U}},\sigma_{\mathrm{U}}) + P(\mathrm{V}|s)\mathcal{N}(f;\mu_s,\sigma_s), \end{aligned} \quad (9.11)$$

where the observation is just the continuous F0 value $\mathbf{o} = f$, $P(\mathrm{U}|s)$ and $P(\mathrm{V}|s)$ are the state-dependent unvoiced or voiced component weights respectively, $P(\mathrm{U}|s) + P(\mathrm{V}|s) = 1$. μ_{u} and σ_{u} are parameters of the globally tied distribution (GTD) for unvoiced speech, and μ_s and σ_s are state-dependent Gaussian parameters for voiced speech. Since the F0 observation is continuous, dynamic features can be easily calculated without considering boundary effects. Consequently, static, delta and delta–delta F0 features are modelled in a single stream using Eq. (9.11).

During HMM training, the initial parameters of the globally tied *unvoiced* Gaussian component can be either pre-defined or estimated on all unvoiced F0 observations. The subsequent training process is similar to standard HMM training. With global tying and random unvoiced F0 observations, the estimated parameters of the unvoiced Gaussian component will have very broad variance and be distinctive from the voiced Gaussian components which model specific modes of the F0 trajectory with much tighter variances. The state-dependent weights of the two components will reflect the voicing condition of each state. During the synthesis stage, similar to MS-DHMM, the weight of the voiced component is compared to a predefined threshold

[4] As implicit voicing condition modelling requires distinct statistical properties between voiced and unvoiced distributions, the interpolation approach in Sect. 9.3.1 is not appropriate here.

to determine the voicing condition. Then the parameters of the voiced Gaussians are used to generate an F0 trajectory for voiced regions as in MSDHMM. For unvoiced states, no F0 values are generated and instead white noise is used for excitation of the synthesis filter.

With the continuous F0 assumption, the limitations of MSDHMM in Sect. 9.2.2 are effectively addressed. Since there is only one single F0 stream, there are no redundant voicing probability parameters. When using the MDL criterion in state clustering (Shinoda and Watanabe 1997) , the removal of redundancy will lead to more clustered states which may model richer F0 variations. More importantly, compared to MSDHMM, the use of a single stream introduces a stronger constraint on the temporal correlation of the continuous F0 observations and this will lead to the generation of more accurate F0 trajectories. It is also worth noting that the use of GTD not only contributes to voicing classification, it has an additional advantage. During HMM training, due to the use of multiple (two) Gaussian components, F0 observations within voiced regions are no longer exclusively assigned to voiced Gaussians. F0 extraction errors may be subsumed by the "background" GTD. This will lead to more robust estimation of the voiced Gaussian parameters than MSDHMM.

9.3.2.2 Continuous F0 Modelling with Independent Voicing Label and F0 Value

To improve the voicing classification performance, voicing labels can be assumed to be observable (Yu and Young 2011). 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 Fig. 9.2c. The state output distribution at state s is then defined as

$$p(\mathbf{o}|s) = p(l, f|s) = p(f|s)^{\gamma_f} P(l|s)^{\gamma_l}, \tag{9.12}$$

where the observation $\mathbf{o} = [f \;\; l]$, $p(f|s)$ and $P(l|s)$ are the distributions for the continuous F0 and voicing label streams respectively, γ_f and γ_l are stream weights. γ_f is set to be 1 and γ_l is set to be a very small positive value ε [5].

Since it is real valued, f is augmented by dynamic features, as in the implicit voicing case. No dynamic features are required for the voicing label l. In CF-IVF, the two streams share the same state clustering structure. Using (9.12), standard maximum likelihood HMM training can be used to estimate parameters of $p(f|s)$ and $P(l|s)$. During synthesis, state voicing status is only determined by the voicing label stream. Each state s is classified as voiced if $P(\mathbb{V}|s)$ is greater than a predefined threshold and unvoiced otherwise. The F0 trajectory is then generated using the same approach as in section MSDHMM.

[5] This means, in HMM training, the voicing labels do not contribute to the forward–backward state alignment stage but their model parameters are updated once the state alignment has been determined.

Since the voicing condition is modelled by an independent data stream, there is no requirement for the statistical properties of the voiced and unvoiced regions to be distinct. Hence, for example, SPLINE interpolation could be used in unvoiced regions in the hope that its tighter variance might lead to better trajectory modelling in V/U boundary regions (Lyche and Schumaker 1973).

In Eq. (9.12), the continuous F0 density $p(f|s)$ can have any form, for example, a single Gaussian. However, even though voicing classification is now explicit, it is still better to use the GTD model defined by (9.11) since the globally tied distribution will absorb F0 estimation errors and therefore provide more robust modelling.

9.3.2.3 Continuous F0 Modelling with Joint Voicing Label and F0 Value

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. Here is a refined approach, 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 Fig. 9.2d. The state output distribution is

$$p(\mathbf{o}|s) = p(l, f|s) = P(l|s)\, p(f|s, l) \tag{9.13}$$

Compared to CF-IVF, CF-JVF introduces correlation between voicing labels l and continuous F0 values f and allows voicing labels to affect the forward–backward state alignment process. This will naturally strengthen the voicing label modelling. It is interesting to see that the DBN of CF-JVF is the same as MSDHMM. However, observation definition is different. In MSDHMM, each observation dimension is a discontinuous variable as defined in Eq. (9.1). In contrast, CF-JVF uses different data types for different dimensions. Each dimension is either discrete or continuous, but not mixed. Only continuous F0 dimensions require calculation of dynamic features.

It can be shown that the update formula for the parameters of $p(f|s, \mathtt{V})$ is the same as the standard ML update formula except for changing the form of state occupancy calculation (Yu and Young 2011). 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 synthesis stage, each state of the HMMs is classified as voiced or unvoiced state by comparing $P(l|s)$ to a predefined threshold.

9.4 Experimental Comparison Between MSDHMM and CF-HMM

The continuous F0 modelling techniques described above have been compared to MSDHMM on two CMU ARCTIC speech synthesis data sets (Kominek and Black 2003). A US female English speaker, `slt`, and a Canadian male speaker, `jmk`, were used. Each data set contains recordings of the same 1132 phonetically balanced sentences totalling about 0.95 h 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 HMM speech synthesis toolkit version 2.0.1 (HMM-based Speech Synthesis System (HTS)). Mixed excitation using STRAIGHT was employed in which the conventional single pulse train excitation for voiced frames is replaced by a weighted sum of white noise and a pulse train with phase manipulation for different frequency bands. The weights are determined based on *aperiodic component* features of each frequency-band (Kawahara et al. 2001). This mixed excitation model has been shown to give significant improvements in the quality of the synthesized speech (Yoshimura 2002).

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). All features were extracted using the STRAIGHT programme (Kawahara et al. 1999a). Spectral, F0 and aperiodic component features were modelled in separate streams during context-dependent HMM training.

For MSDHMM, as indicated in Sect. 9.2.2, separate streams have to be used to model each of the static, delta and delta–delta F0 features (Masuko et al. 2000). In contrast, all CF-HMM systems used a single stream for static and dynamic features of the continuous F0 observations. The CF-HMM with explicit voicing condition modelling also had an extra data stream for voicing labels. During HMM training for all systems, the stream weight for the aperiodic components was set to to be a very small positive value ϵ, similar to that of the voicing label in Sect. 9.3.2.2. MDL-based state clustering (Shinoda and Watanabe 1997) 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 similar number of parameters as the continuous F0 modelling techniques. The same MDL factor is used for comparing CF-IVF and CF-JVF. The duration of each HMM state is modelled by a single Gaussian distribution (Yoshimura et al. 1998). A separate state clustering process was then performed for the duration model parameters. During the synthesis stage, global variance (GV) was used in the speech parameter generation algorithm to reduce the well-known over-smoothing problem of HMM based speech synthesis (Toda and Tokuda 2007).

Figure 9.3 shows an example of the F0 trajectories generated by the two models compared to natural speech. Similar trends as shown by the objective measures can be observed: the CF-HMM F0 trajectory is a closer match to the natural speech whilst the MSDHMM has more accurate voicing classification. When listening to the speech, it can be perceived that both the natural speech and CF-HMM synthesised speech have a distinct rise at the end, whilst the MSDHMM speech was flat. In contrast, the effect of the voicing classification errors was not perceptible. Quantative comparisons are given as below.

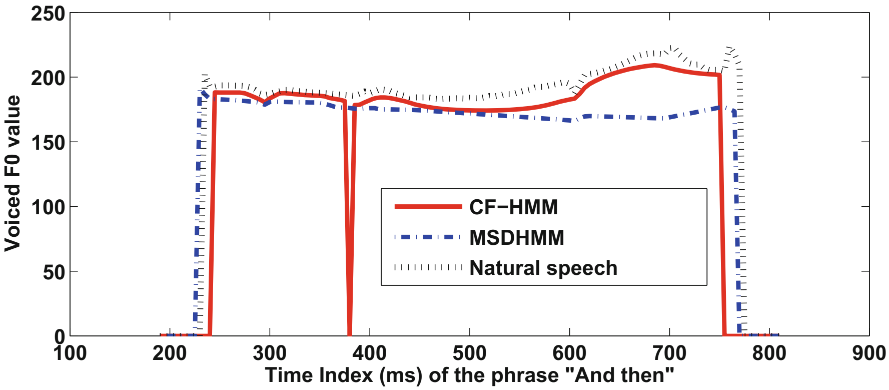


Fig. 9.3 Example F0 trajectories generated by the MSDHMM and CF-HMM models compared to natural speech

9.4.1 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 both the MSDHMM and CF-HMM systems. To reduce the effect of the duration model when comparing the generated F0 trajectories, state level durations were first obtained by forced-aligning the known natural speech from the test set. Then, given the natural speech durations, 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. The root mean square error of F0 is defined as

$$\mathtt{RMSE} = \sqrt{\frac{\sum_{t\in\mathcal{V}} (f(t) - f_{\mathtt{r}}(t))^2}{\#\mathcal{V}}}, \tag{9.14}$$

where $f_{\mathtt{r}}(t)$ is the extracted F0 observation of natural speech at time t, $f(t)$ is the synthesized F0 value at time t, $\mathcal{V} = \{t : l(t) = l_{\mathtt{r}}(t) = \mathtt{V}\}$ 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 rate of mismatched voicing labels

$$\mathtt{VCE} = 100\frac{\sum_{t=1,T} \delta(l(t), l_{\mathtt{r}}(t))}{T} \tag{9.15}$$

where $\delta(l, l_r)$ is 1 if $l = l_r$ and 0 otherwise, and T is the total number of frames.

From Table 9.1, CF-HMM approaches effectively reduce the average F0 synthesis errors (RMSE) in both training and test sets compared to MSDHMM. This demonstrates the effectiveness of using continuous F0 observations. On the other hand, VCE performance becomes worse when continuous F0 assumption is used.

Table 9.1 Objective comparisons between MSDHMM and CF-HMM approaches

Data set	F0 modelling	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

CF-GTD has the worst performance due to weak modelling of voicing labels. By explicitly modelling observable voicing labels, CF-IVF obtains significant improvement. CF-JVF can achieve further improvement on VCE due to the strengthened correlation between voicing label and continuous F0 values. CF-JVF also improves the RMSE, i.e. F0 trajectory modelling, of most test sets except for the female test set. However, the VCEs of CF-HMM are still worse than MSDHMM. This is expected since MSDHMM assumes observable voicing labels and dependency between F0 observations, which leads to stronger voicing condition modelling.

9.4.2 Subjective Listening Tests

To properly measure performance of the synthesis systems, two forms of subjective listening tests were conducted. The CF-IVF system was first used as the representative approach of CF-HMM and then was compared with other CF-HMM approaches.

First, a mean opinion score (MOS) test was conducted. Thirty sentences were selected from the held-out test sets and each listener was presented with ten sentences randomly selected from them, of which five were male voices and the other five were female. The listener was then asked to give a rating from 1 to 5 to each utterance. The definition of the rating was: 1-bad, 2-poor, 3-fair, 4-good, 5-excellent. In total, 12 non-native and 11 native speakers participated in this test. In order to focus the evaluation on F0 synthesis, the state durations were obtained by forced-aligning the natural speech with known phone context transcriptions. Also, the spectral and aperiodic component features used were extracted from natural speech. Thus, the CF-HMM and MSDHMM models were only used to perform voicing classification of each state and generate F0 trajectories for the voiced regions. In addition, vocoded

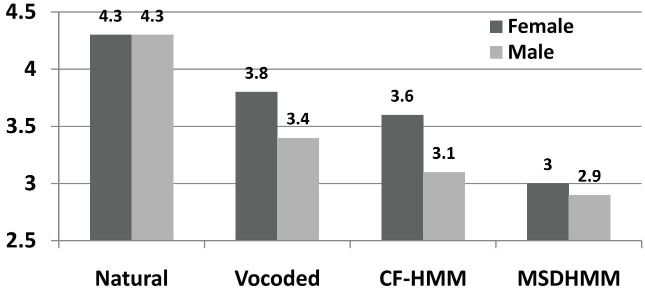


Fig. 9.4 Mean opinion score comparison of CF-HMM (CF-IVF) vs MSDHMM for F0 modelling (spectral, aperiodic component and durational features are identical across all systems). Also included for comparison are the MOS scores for natural and vocoded speech

speech [6] and natural speech were also included in the test to determine the effects of vocoder artifacts on the assessment.

Figure 9.4 shows the resulting MOS scores. It can be observed that the CF-HMM system outperformed the MSDHMM system for both male and female speakers. Vocoded speech, which may be regarded as the best possible speech that could be synthesised from any statistical model, was better than speech synthesized using either the CF-HMM or MSDHMM systems. However, the degradation from natural speech to vocoded speech was much larger than the degradation from vocoded speech to CF-HMM synthesised speech. It can also be observed that speech quality degradation of the female speaker is less than that of the male speaker. Pair-wise two-tail Student's t-tests were performed to evaluate the statistical difference between different systems. With a 95% confidence level, CF-HMM was significantly better than MSDHMM for the female speaker ($p = 0.004$), while the gain for the male speaker was not statistically significant ($p = 0.18$). This suggests that male speech is less sensitive to continuous F0 modelling. The vocoded speech was not significantly different from CF-HMM for both speakers (female: $p = 0.20$, male: $p = 0.08$). Thus, as far as statistical F0 modelling is concerned, on this data, the CF-HMM system is comparable in naturalness to vocoded speech [7].

The above MOS test used ideal duration, spectral and aperiodic component features. To compare the actual performance of complete synthesis systems, pair-wise preference tests were conducted. 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

[6] Vocoded speech is the speech synthesized from the original spectral, F0 and aperiodic component features of natural speech. The only loss during this process comes from feature extraction and synthesis filter.

[7] Although there was no significant difference between CF-HMM and MSDHMM for the male speaker, the t-test showed that vocoded speech was significantly better than MSDHMM for both speakers (female: $p = 0.00005$, male: $p = 0.005$).

Fig. 9.5 Comparison between CF-HMM (CF-IVF) and MSDHMM

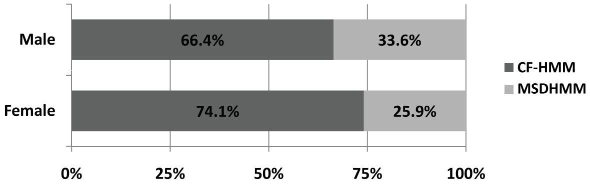
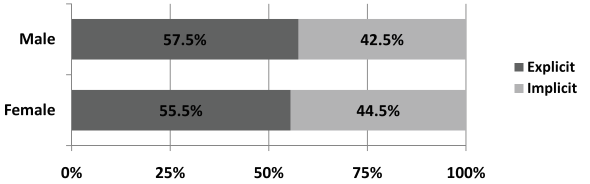


Fig. 9.6 Comparison between implicit and explicit voicing condition modelling



useful test of the generalization ability 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 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. Altogether 12 native and 10 non-native speakers participated in the test. The result is shown in Fig. 9.5.

It can be observed that the CF-HMM system outperformed the MSDHMM system for both male and female speakers. Statistical significance tests were also performed assuming a binomial distribution for each choice. The preference for CF-HMM was shown to be significant at 95 % confidence level (p-values for both speakers are approximately 0). Similar to the MOS test, the CF-HMM was also more dominant for the female speaker than the male speaker.

The above CF-HMM system is a CF-IVF system with 1-best selection approach for unvoiced F0 generation. It is also interesting to compare different CF-HMM approaches. First, CF-GTD is compared to CF-IVF, which is also the comparison between *implicit* and *explicit* voicing condition modelling. A panel of 21 subjects (10 non-native and 11 native speakers) was used.

As can be seen in Fig. 9.6, explicit modelling is better than implicit modelling for both speakers. This is consistent with the objective comparison in Table 9.1. Statistical significance tests showed that the difference was significant for the male speaker ($p = 0.01$) and almost significant for the female speaker ($p = 0.05$).

When explicit voicing condition modelling is used, the dependency between voicing label and F0 observation can be reserved, which leads to CF-JVF. The two approaches are compared in Fig. 9.7. It can be observed that the CF-JVF system outperformed the CF-IVF system for both male and female speakers. Statistical significance tests were performed for the result assuming a binomial distribution for

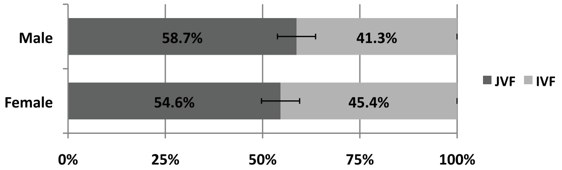


Fig. 9.7 Comparison between CF-IVF and CF-JVF. Confidence interval of 95 % is shown

each choice. The preference for CF-JVF was shown to be significant at 95 % confidence level (p-values: 0.03 for female and 0.0002 for male). This is also consistent with the objective measures.

In summary, the CF-HMM framework addresses many of the limitations of the most widely used DF-HMM approach, MSDHMM. It has been shown to yield improved performance. It is also more consistent with HMM-based speech recognition systems and it can, therefore, more easily share existing techniques, algorithms and code.

9.5 Further Analysis

Although the CF-HMM has been shown to yield significant improvement in speech quality compared to the MSDHMM in the previous section, 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 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.

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 (Zen et al. 2001), 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, Eq. (9.6), GTD can be easily introduced. Assuming the F0 extraction error is independent of states and combining the prior weights together, Eq. (9.6) becomes

$$p(f_+|s) = c_1^s\delta(f_+,\text{NULL}) + \big(c_2^s\mathcal{N}(f|\text{U}) + c_3^s\mathcal{N}(f|s,\text{V})\big)(1-\delta(f_+,\text{NULL})) \tag{9.16}$$

and $c_1^s = c_{\text{U}}^s\lambda_{\text{U}}^d + c_{\text{V}}^s\lambda_{\text{V}}^d$, $c_2^s = c_{\text{U}}^s\lambda_{\text{U}}^c$, $c_3^s = c_{\text{V}}^s\lambda_{\text{V}}^c$, $c_1^s + c_2^s + c_3^s = 1$.

Table 9.2 Objective comparison between MSDHMM extensions and CF-HMM (CF-IVF)

Data set	F0 modelling	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-IVF	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-IVF	12.58	7.29	11.90	8.43

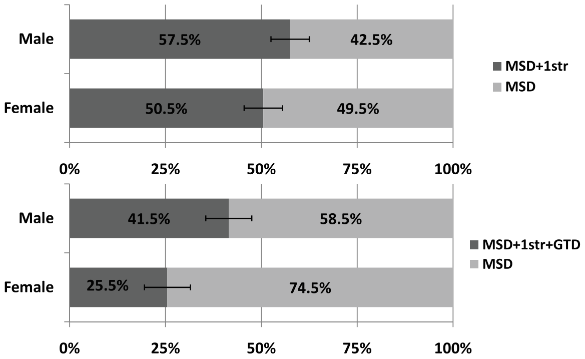
Given that a single F0 stream and GTD can both be implemented within the DF-HMM framework, the MSDHMM can be extended to include these and thereby, allow a direct comparison with the CF-HMM. To use a single F0 stream, SPLINE interpolation is first performed for all unvoiced segments and dynamic real-valued 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 $\mathcal{N}(f|\mathtt{U})$ in Eq. (9.16) 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[8]. During synthesis, c_1^s is compared to a pre-determined threshold (0.5 in this chapter) to determine the voicing classification for each synthesis frame.

Experiments comparing the extended MSDHMM systems and the CF-HMM system were performed to demonstrate the above. Data and experimental set up are the same as in Sect. 9.4. Again, CF-IVF is used as CF-HMM in the experiments.

From the objective comparison Table 9.2, 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

[8] 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.

Fig. 9.8 MSDHMM vs. extended MSDHMM. Confidence interval of 95% is shown

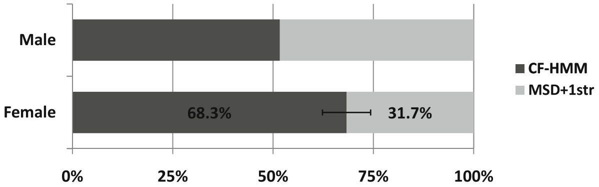


Gaussian component because it can absorb mass from real-valued F0 observations in voiced regions. In contrast to the MSDHMM, the CF-HMM has randomly generated unvoiced F0 values which provide a strong statistical constraint (especially in the dynamic features) that prevents the GTD component from subsuming the correctly estimated voiced F0 observations. Hence, although the GTD can absorb F0 outliers and yield robust F0 estimation in the CF-HMM, it cannot do the same for the MSDHMM (Yu and Young 2011). It is worth noting that from the definition of RMSE, Eq. (9.14), 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 DF-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.

Figure 9.8 shows the comparison between the two extended MSDHMM systems and the traditional MSDHMM. Eight 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. 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 9.9 shows the comparison between CF-HMM and MSDHMM with a single F0 stream, which outperformed the traditional MSDHMM. Eight 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

Fig. 9.9 CF-HMM vs. MSDHMM with single F0 stream. Confidence Interval of 95 % is shown



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 Fig. 9.9 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.

Acknowledgement Most of the above works were done at the Machine Intelligence Lab, Cambridge University Engineering Department, during the author's tenure. The author would like to thank all his colleagues, especially Steve Young, Blaise Thomson and Tomoki Toda, at Cambridge University for their valuable contributions. The work was also supported by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning and the China NSFC project No. 61222208.

Appendix

The definition of $p(f_+)$ follows the standard approach for distributions of mixed type (discrete and continuous). (Papoulis 1984) 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 (Rudin 1987).

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 (Rudin 1987)). 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:

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

where $\lambda^d + \lambda^c = 1$. Note that the probability function has only been defined in terms of open intervals and the $\{\mathtt{NULL}\}$ 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 $f_+ \in \{\mathtt{NULL}\} \cup (-\infty, \infty)$ by:

$$p(f_+) = \lambda^d\,\delta(f_+, \mathtt{NULL}) + \lambda^c\,\mathcal{N}(f)(1 - \delta(f_+, \mathtt{NULL})). \tag{9.17}$$

This form of density function can be used in likelihood calculation 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 $a, b \in \mathbb{R}, a < b$):

$$\mu(\{\mathtt{NULL}\}) = 1, \tag{9.18}$$

$$\mu((a,b)) = (b - a). \tag{9.19}$$

Using Lebesgue integration (Rudin 1987) of the probability density p, Eq. (9.17), with respect to this measure gives that:

$$P(A) = \int_A p\, d\mu. \tag{9.20}$$

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

$$P(\{\mathtt{NULL}\}) = p(\mathtt{NULL}) = \lambda^d, \tag{9.21}$$

$$P((a,b)) = \int_{f\in(a,b)} p(f)df, \tag{9.22}$$

$$= \lambda^c \int_{f\in(a,b)} \mathcal{N}(f)df. \tag{9.23}$$

References

Freij, G. J., and F. Fallside. 1988. Lexical stress recognition using hidden Markov modeld. In *ICASSP*, 135–138.

HMM-based speech synthesis system (HTS) 2007. http://hts.sp.nitech.ac.jp. Accessed on 1 July, 2008.

Huang, X., A. Acero, and H. Hon. 2001. *Spoken Language Processing*. Upper Saddle River: Prentice Hall PTR.

Imai, S. 1983. Cepstral analysis synthesis on the mel frequency scale. In *Proceedings of ICASSP*, 93–96.

Kawahara, H., I. M. Katsuse, and A.D. Cheveigne.1999a. 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* 27 (3–4): 187–207.

Kawahara, H., H. Katayose, A. D. Cheveigne, and R. D. Patterson. 1999b. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. In *Proceedings of EUROSPEECH*, 2781–2784.

Kawahara, H., J. Estill, and O. Fujimura. 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *Proceedings of MAVEBA*, Firentze, Italy, 13–15.

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

Lyche, T., and L. L., Schumaker. 1973. On the convergence of cubic interpolating splines. *Spline functions and approximation theory*. Birkhauser, 169–189.

Masuko, T., K. Tokuda, N. Miyazaki, and T. Kobayashi. 2000. Pitch pattern generation using multi-space probability distribution HMM. *IEICE Transaction* J83-D-II (7): 1600–1609.

Papoulis, A. 1984. *Probability, random rariables, and stochastic processes*. U.S.: McGraw-Hill.

Ross, K. N., and M. Ostendorf. 1999. A dynamical system model for generating fundamental frequency for speech synthesis. *IEEE Transactions on Speech and Audio Processing* 7 (3): 295–309.

Rudin, W. 1987. *Real and complex analysis*, 3rd ed. New York: McGraw-Hill.

Shinoda, K., and T. Watanabe. 1997. Acoustic modeling based on the MDL principle for speech recognition. In *Proceedings of EUROSPEECH*, 99–102.

Talkin, D. 1995. A robust algorithm for pitch tracking (RAPT). In *Speech coding and synthesis*, ed. W. B. Kleijn and K. K. Paliwal, 497–516. Amsterdam: Elsevier.

Toda, T., and K. Tokuda. 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions on Information and Systems* E90-D (5): 816–824.

Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proceedings of ICASSP*, 1315–1318.

Tokuda, K., T. Mausko, N. Miyazaki, and T. Kobayashi. 2002. Multi-space probability distribution HMM. *IEICE Transactions on Information and Systems* E85-D (3): 455–464.

Yoshimura, T. 2002. Simultaneous modelling of phonetic and prosodic parameters, and characteristic conversion for HMM based text-to-speech systems, PhD diss, Nagoya Institute of Technology.

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 1998. Duration modelling in HMM-based speech synthesis system. In *Proceedings of ICSLP*, 29–32.

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 1999. Simultaneous modelling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of EUROSPEECH*, 2347–2350.

Yu, K., and S. Young. 2011a. Continuous F0 modelling for HMM based statistical speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing* 19 (5): 1071–1079.

Yu, K., and S. Young. 2011b. Joint modelling of voicing label and continuous f0 for hmm based speech synthesis. In *Proceedings of ICASSP*.

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

Yu, K., B. Thomson, and S. Young.2010. From discontinuous to continuous F0 modelling in HMM-based speech synthesis. In *Proceedings of ISCA SSW7*.

Zen, H., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 2001. A pitch pattern modeling technique using dynamic features on the border of voiced and unvoiced segments. *Technical report of IEICE* 101 (325): 53–58.