Using Noisy Speech to Study the Robustness of a Continuous F0 Modelling Method in HMM-based Speech Synthesis

Kalu U. Ogbureke, João P. Cabral, Julie Carson-Berndsen

CNGL, School of Computer Science and Informatics, University College Dublin,
Belfield, Dublin 4, Ireland

kalu@ucdconnect.ie, joao.cabral@ucd.ie, julie.berndsen@ucd.ie

Abstract

In parametric text-to-speech synthesis using Hidden Markov Model (HMM), the fundamental frequency (F0) parameter modelling is important because it has a direct effect on the prosody of synthetic speech. F0 is typically modelled by a discrete distribution for unvoiced speech and a continuous distribution for voiced, by using a multi-space distribution (MSD). However, F0 modelling using MSD-HMM is not accurate around the voiced-unvoiced (V-UV) and (UV-V) transitions and it is affected by voicing decision errors of the F0 estimation algorithm. In order to reduce this problem, HMM-based speech synthesisers have been proposed that model F0 using continuous HMM. This approach usually obtains the continuous F0 contours by interpolating F0 in unvoiced regions. The problem with this method is that it is affected by voiced decision errors during speech analysis. For example, if voiced speech segments are incorrectly classified as unvoiced, the F0 contour in this region is obtained by interpolation which might be a poor estimate of the natural F0. This paper proposes to use an F0 estimation method that does not require a hard voiced/unvoiced classification and produces a reasonable smooth F0 contour. The robustness of this method was studied in the conditions of high-quality recorded speech and recorded speech with additive noise. The motivation for using noisy speech was to study the effect of voiced decision errors on the quality of the synthetic speech.

Index Terms: continuous F0 modelling, voicing strength, HMM-based speech synthesis

1. Introduction

Accurate prediction of F0 in speech synthesis is important because F0 carries prosodic information and it affects the perceptual quality of synthetic speech. In HMM-based speech synthesis, contribution to F0 errors come from the F0 estimation method as well as the weakness of the underlying statistical method.

The standard method for modelling F0 in HMM-based speech synthesis is to use MSD-HMM [1]. This extension to HMM is used because the estimated F0 contour is assumed to be made of continuous values in voiced regions and discrete values in unvoiced regions (set equal to zero). MSD-HMM permits to model F0 in voiced regions using a continuous probability distribution and using a discrete distribution in unvoiced regions. However, the voiced/unvoiced (V/U) decision, during synthesis, is based on weights of distributions in each state which is prone to error when the difference between weights of distributions is small [2].

In order to overcome the limitations of MSD-HMM for F0 modelling in statistical speech synthesis, recent improvements have been proposed which are based on statistical models using continuous distributions only. One such improvement consists of using globally tied distribution HMM (GTD-HMM) [3] instead of MSD-HMM. The advantage of GTD-HMM over MSD-HMM is better modelling of F0 at the transitions between voiced and unvoiced regions, e.g., dynamic features are estimated at the boundaries between voiced and unvoiced regions unlike in MSD-HMM. However, in both methods the V/U decision during synthesis is state-based and prone to error. An alternative approach is to obtain a continuous F0 contour during speech analysis and to model F0 using continuous HMM. This technique has been used by interpolating F0 in unvoiced regions using a spline or cubic function, e.g. [4], or by using a mixture of random noise and an exponential decay function of a running average, e.g. [5]. In this case, the estimation and continuous modelling of an additional parameter called voicing strength is necessary, for deciding if a speech frame is voiced or unvoiced during synthesis. However, a weakness of this method is that it is affected by the V/U decision of the F0 estimation method during analysis. For instance, the interpolated F0 values might be a poor estimation of the natural F0 values for frames incorrectly classified as unvoiced.

This paper proposes to estimate continuous F0 contour directly from the speech using a glottal closure instant (GCI) detection method. The advantage of this approach is that it avoids the effects of V/U decision errors in continuous F0 contours estimated during analysis. In order to better study this effect of V/UV decision errors on the quality of the synthetic speech, a noisy speech database was used in the experiments. Another motivation for using noisy speech is that clean speech is not always available in certain real-time applications of HMM-based speech synthesis. For example, there are techniques to adapt HMMs to the speaker's voice using a small amount of data from the target speaker. However, this speech data might be noisy if the speaker is using low quality recording equipment such as a mobile device or if the recording environment is noisy.

2. HMM-based Speech Synthesiser Using Continuous F0

2.1. F0 Estimation Based on GCI Detection

The F0 of a voiced speech segment can be determined using the GCIs, which are also called instants of maximum excitation or epochs. F0 is calculated as $F0 = 1/T0$, in which the fundamental period T0 is estimated as the duration between consecutive GCIs. In this work, the GCIs are obtained using a method called SEDREAMS (Speech Event Detection using the Residual Excitation And a Mean-based Signal) [6]. This method is a two-step process. In the first step, the time interval between which the GCI is expected to lie is estimated as the interval between the minima and following positive zero-crossing of a mean-based signal, which is calculated from the speech signal. In the second step, the GCI is estimated by finding a local maximum in the LP-residual that falls within the expected interval.

The SEDREAMS method was chosen in this work because it is accurate, robust to noise and also detects GCIs in regions of unvoiced speech. Note that the GCI is meaningless for unvoiced speech because there is no glottal activity during human production of unvoiced speech. The GCIs detected in voiceless regions correspond to peaks located at unpredictable points within the analysis interval and that depend on the SEDREAMS algorithm constraints (such as minimum and maximum F0). However, such method is expected to correctly capture GCIs in regions which are a mix of voiced and voiced speech, such as voiced fricatives, which are often not detected by most epoch detection algorithms which perform a voiced/unvoiced classification.

Finally, in this work a smoothing operation using the median function is performed on the resulting F0 contour in unvoiced regions, to avoid peaks in the F0 contour which might affect the statistical modelling by HMMs.

2.2. Voicing Strength Estimation Method

2.2.1. Voicing Strength Estimation from Residual Harmonics

The V/U decision method used in this work is based on the information in the residual harmonics [7]. The first step in the estimation of the residual harmonics is the estimation of the spectral envelope from the speech signal $S(t)$ by auto-regressive modelling. This is followed by estimation of the residual signal $e(t)$ by inverse filtering. Then, the residual signal $e(t)$ is segmented using a Hanning window and the amplitude spectrum $E(f)$ of the short-time signal is calculated. The summation of the residual harmonics is computed in the frequency range between the minimum and the maximum F0 from the amplitude spectrum $E(f)$ as follows:

$$SRH(f) = E(f) + \sum_{k=2}^{K} \left(E(k.f) - E((k-\frac{1}{2}).f) \right), \quad (1)$$

where K is the number of harmonics in the frequency interval. K is set equal to 5 in this experiment. The frequency maximising $SRH(f)$ is an estimate of the fundamental frequency, $\hat{F}0$, and a frame is classified as voiced if $SRH(\hat{F}0)$ is greater than a fixed threshold (equal to 0.07).

In this work, the voicing strength parameter v_i is calculated for each speech frame i by performing a normalisation of the parameter $SRH(\hat{F}0)$, since the voicing strength is usually defined between 0 and 1 (represents the probability of voicing). The normalisation is performed as follows:

$$v_i = \frac{SRH_i(\hat{F}0)}{SRH^{max}(\hat{F}0)}, \quad (2)$$

where SRH^{max} is the maximum of $SRH_i(\hat{F}0)$ over all frames.

2.2.2. Calculation of V/U Threshold

An experiment was conducted to obtain the voicing strength threshold for V/UV classification. First, the reference V/UV classification was performed using the method described in [6] on the utterances of the RMS voice of the CMU_ARCTIC corpus [8] of read speech. Then, the V/UV classification was performed using the voicing strength threshold for different values of this threshold. The optimal voicing strength threshold was the one that minimised the voicing decision errors. A voiced error occurs when a reference voiced frame is classified as unvoiced and an unvoiced error occurs when a reference unvoiced frame is classified as voiced.

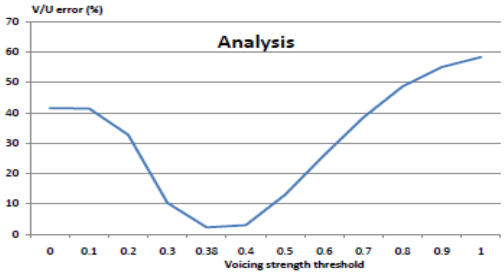

Figure 1: Variation of V/U error rate with voicing strength for clean speech data.

Figure 1 shows the variation of the V/U error rate (number of V/UV frame errors normalised by the total number of speech frames) as a function of the voicing strength threshold. The minimum error rate occurs at 0.38.

2.3. Speech Analysis

The speech parameters estimated during the analysis part of the HMM-based speech synthesiser are F0, voicing strength, spectrum and aperiodicity parameters, with their delta and delta-delta features. The STRAIGHT method [9] is used to calculate the spectral envelope and aperiodicity measurements, which are converted to 24^{th} order MGC and the mean values of the aperiodicity measurements over five frequency bands respectively.

2.4. Statistical Modelling

The statistical modelling is performed using the HTS toolkit version 2.1 [10]. The model topology is a five-state left-to-right Hidden semi-Markov model (HSMM) with four streams. HSMM is an extension of HMM for modelling speech duration explicitly. Each stream is clustered and tied using different decision trees because the speech parameters are assumed to be independent.

2.5. Speech Synthesis

During synthesis, speech parameters are generated from HSMM using the maximum likelihood criterion. The V/U decision on each generated frame is based on the optimal threshold obtained during analysis (see section 2.2.2) and the generated voicing strength parameter. Finally, the speech waveform is generated from the parameters using the STRAIGHT vocoder [9].

3. Experimental Results

3.1. Systems

Two HMM-based speech synthesisers using continuous F0 modelling were compared in the experiments. One is the proposed HMM-based speech synthesiser which uses a GCI detection method to extract continuous F0 contours during analysis. The other is a baseline system which is similar to the proposed system, but it performs a V/U decision during analysis (using the method proposed by [7] which was described in Section 2.2.1). The F0 values estimated by the baseline system are equal to those of the proposed system in the voiced regions (calculated using the same GCI detection method). However, the F0 values of the baseline in the unvoiced regions are obtained from the voiced F0 values by interpolation using a cubic function. The statistical modelling and speech synthesis parts were similar for both systems.

There were three versions of the baseline system that differed in the voicing threshold used to make the voicing decision during analysis. The first system used the optimal voicing threshold (0.38), the second used a lower voicing threshold (0.3) and the third used a higher voicing threshold (0.43). The three thresholds were chosen to study the effect of voicing decision errors on F0 modelling. By increasing the voicing threshold there is an increase in the number of voiced frames classified as unvoiced and consequently the negative effect of F0 interpolation in the baseline system is more prominent.

3.2. Speech Corpus

Two types of speech data were used in the experiments, the clean and noisy speech. The clean speech was the RMS voice of CMU_ARCTIC corpus [8] of read speech that contains a total of 1132 sentences. Meanwhile, the noisy speech was obtained by adding white noise to the clean speech at 0 and 3 dB. Both speech datasets were divided into a training set composed of 1030 sentences and a test set (the remaining 102 sentences).

The speech dataset with additive noise was used to study the performance of the HMM-based speech synthesisers trained on noisy speech data (characterised by higher voiced error rate). It is expected that the difference in synthesised speech quality between the two systems is more significant for noisy speech than for clean speech, because the amount of voicing classification errors during analysis is expected to increase with the amount of additive noise for the baseline system, whereas the proposed system is not affected by this problem.

Figure 2 shows an example of the F0 contour estimated using the glottal epoch detection method and the interpolated F0 contour estimated using the baseline system, for an utterance of the noisy speech dataset. The differences between the contours in the voiced regions (represented by ’V’), which were obtained from clean speech, suggest that voicing classification errors in the baseline system will negatively affect F0 modelling. For example, the difference is particularly significant around the frame number 200.

3.3. Synthetic Speech

The values of the voicing strength, spectral and aperiodicity parameters generated by the two HMM-based speech synthesisers were equal, since the only difference between the systems was the method to estimate F0. The spectral and aperiodicity parameters values used for generating the speech waveforms were the same as those estimated for clean speech, in order to study the effect of noise on F0 modelling only. Another reason for this decision was to avoid the masking effect of noise on the perception of pitch by the listeners during the experiments.

The speech parameters were generated by HSMMs from the test sentences by imposing the natural durations of the utterances in the test set. This allowed to compute error measures between the F0 contours generated by the HMM-based speech synthesisers and the original contours computed during analysis. Speech was synthesised using the proposed and baseline systems as described in Section 2.5.

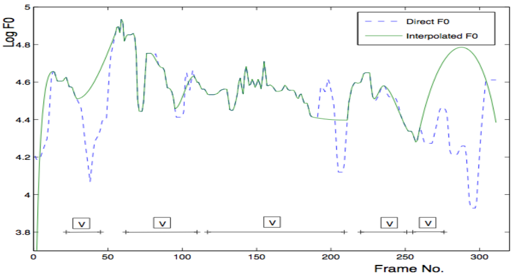

Figure 2: An example of F0 contours by direct and interpolation methods for noisy speech at 3 dB. ’V’ represents voiced regions. Voiced regions were obtained by V/U decision of the clean utterance.

3.4. Objective Evaluation

3.4.1. V/UV Error Measurement

The criterion used for the objective evaluation was the Root Mean Squared Error (RMSE) between the reference F0 estimated using the reference method described in section 2.1 and the F0 generated by the HMM-based speech synthesisers on the test set for voiced frames only.

3.4.2. Results

Table 1 shows the results of the objective evaluation for the four systems trained on clean speech. These results show that the proposed HMM-based speech synthesiser generates F0 values that better approximate the natural F0 values as compared with the systems that use F0 interpolation in unvoiced regions. Furthermore, results show that an increase in the number of voicing decision errors during analysis (due to an increase in the v_i threshold), degrades F0 modelling for the baseline systems.

Table 1: RMSE values obtained for the baseline systems and the proposed system for clean speech.

System (T = Threshold in Analysis)	**RMSE**
T = 0.3 (lower threshold)	0.1128
T = 0.38 (optimal threshold)	0.1140
T = 0.43 (higher threshold)	0.1152
Proposed approach	0.1120

Table 2 shows the results for noisy speech condition at 0 and 3dB respectively (using the optimal v_i threshold). The improvement in F0 modelling by the proposed system compared with the baseline is more significant for the noise condition. This result is explained by an increase in the number of voiced frames incorrectly classified as unvoiced due to the addition of noise, because F0 modelling for these frames is affected by F0 interpolation in the baseline system.

3.5. Subjective Evaluation

3.5.1. Experiment

An ABX choice subjective experiment was conducted to evaluate the perceptual quality of synthetic speech between the baseline and the proposed systems for noisy speech at 3 dB.

Table 2: RMSE for the baseline systems and the proposed approach for noisy speech at 0 and 3 dB.

System	RMSE
0 dB interpolated	0.1548
0 dB proposed	0.1287
3 dB interpolated	0.1536
3 dB proposed	0.1241

For this experiment only the noisy speech data was used because the results of the objective evaluation indicated that the difference between the two systems was more significant for this data set. However, we plan to extend these experiments to clean speech because such study could be important for practical applications in which clean speech is available.

15 sentences were randomly selected from the test set, whereby each pair of utterances (A/B) consisted of the same sentence synthesised by the two systems (i.e. the baseline and proposed systems). The number of participants were 11, made of 5 native and 6 non-native speakers of English, some of whom are speech synthesis experts. They were asked to select the sample (A or B) of each pair that sounded most natural. The third option 'X' was chosen by the subjects when they did not perceive any difference between the two samples.

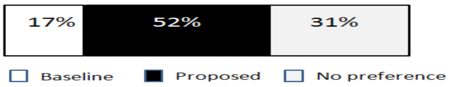

Figure 3: Preference rates between the proposed system and the baseline system for noisy speech at 3 dB.

3.5.2. Results

Figure 3 shows the preference rates obtained for the baseline and the proposed systems, as well as the "no preference" rate. Results show that speech synthesised by the proposed system sounds more natural than speech synthesised by the system based on interpolation, in general. This is because the addition of white noise increased the voicing decision errors in the baseline system which affected more the quality of the baseline system. These results are in agreement with those obtained by objective measurements which indicate that voicing decision errors in the baseline system (which are assumed to increase with addition of noise) degrade F0 modelling.

4. Conclusions

The speech quality of state-of-the-art HMM-based speech synthesisers is affected by voiced decision errors during speech analysis. This paper proposes an approach to overcome this problem which consists of estimating F0 values for voiced and unvoiced speech without making any voiced/unvoiced classification. In this method, F0 is estimated from speech using a technique for glottal closure instant (GCI) detection that allows us to directly obtain continuous and approximately smooth F0 contours. The advantage of this method compared with recently proposed methods which interpolate F0 in unvoiced regions to obtain continuous F0 values is that it does not depend on the performance of a voiced/unvoiced classifier.

The HMM-based speech synthesiser using the F0 extraction technique based on GCI detection was compared against a baseline system which used a F0 interpolation method. The systems were evaluated on clean and noisy speech (as training data). The use of noisy speech was particularly relevant in this study because it permitted a better comparison between the techniques for obtaining continuous F0 contours used by the two systems.

An objective experiment showed that the proposed system produced F0 contours more similar to the natural F0 estimated during analysis for both clean and noisy speech. A perceptual experiment also showed that synthetic speech from the system that used GCIs to estimate F0 sounded perceptually more natural than speech synthesised with the baseline system, in general.

Extended experiments are required for performing a more complete evaluation of the proposed system. For example, an experiment to evaluate the synthetic speech quality of fricatives and words which are mixtures of voiced and unvoiced speech segments as these segments are more likely to contain V/U decision errors.

5. Acknowledgments

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/CE/I1142 as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

6. References

[1] Tokuda, K., Mausko, T., Miyazaki, N. and Kobayashi, T., "Multi-space probability distribution HMM", IEICE Transanction on Information and System, vol.E85-D no.3, 455-464, 2002.

[2] Zhang, Q., Soong, F., Qian, Y., Yan, Z., Pan, J. and Yan, Y., "Improved modeling for F0 generation and V/U decision in HMM-based TTS", In Proc. of ICASSP, 4606-4609, 2010.

[3] Yu, K., Toda, T., Gasic, M., Keizer, S., Mairesse, F., Thomson, B. and Young, S., "Probabilistic modelling of F0 in unvoiced regions in HMM-based speech synthesis", In Proc. of ICASSP, 19-24, 2009.

[4] Kunikoshi, A., Qian, Y., Soong, F. and Minematsu, N., "Improved F0 modeling and generation in voice conversion", In Proc. of ICASSP, 4568-4571, 2011.

[5] Chen, C . J., Gopinath, R. A., Monkowski, M. D., Picheny, M. A. and Shen, K., "New methods in continuous Mandarin speech recognition", In Proc. of Eurospeech, 1543-1546, 1997.

[6] Drugman, T. and Dutoit, T., "Glottal closure and opening instant detection from speech signals", In Proc. of Interspeech, 28912894, 2009.

[7] Drugman, T. and Alwan, A., "Joint Robust voicing detection and pitch estimation based on residual harmonics", In Proc. of Interspeech, 1973-1976, 2011.

[8] Kominek, J. and Black, A., "The CMU Arctic speech databases", In Proc. of 5th ISCA speech synthesis workshop, 223-224, 2004.

[9] Kawahara, H., Masuda-Katsuse, I. and Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds", Speech Communication, Vol. 27, pp. 187–207, 1999.

[10] Drugman, T., "GLOttal Analysis Toolbox", tcts.fpms.ac.be/drugman/Toolbox/, 2011.

[10] "HMM-based speech synthesis system version 2.1", http://hts.sp.nitech.ac.jp, 2008.