

# Parameterization of Vocal Fry in HMM-Based Speech Synthesis

Hanna Silén<sup>1</sup>, Elina Helander<sup>1</sup>, Jani Nurminen<sup>2</sup>, Moncef Gabbouj<sup>1</sup>

<sup>1</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland

<sup>2</sup>Nokia Devices R&D, Tampere, Finland

hanna.silen@tut.fi, elina.helander@tut.fi, Jani.K.Nurminen@nokia.com

## Abstract

HMM-based speech synthesis offers a way to generate speech with different voice qualities. However, sometimes databases contain certain inherent voice qualities that need to be parametrized properly. One example of this is vocal fry typically occurring at the end of utterances. A popular mixed excitation vocoder for HMM-based speech synthesis is STRAIGHT. The standard STRAIGHT is optimized for modal voices and may not produce high quality with other voice types. Fortunately, due to the flexibility of STRAIGHT, different F0 and aperiodicity measures can be used in the synthesis without any inherent degradations in speech quality. We have replaced the STRAIGHT excitation with a representation based on a robust F0 measure and a carefully determined two-band voicing. According to our analysis-synthesis experiments, the new parameterization can improve the speech quality. In HMM-based speech synthesis, the quality is significantly improved especially due to the better modeling of vocal fry.

**Index Terms:** speech synthesis, hidden Markov models, vocal fry, mixed excitation, STRAIGHT

## 1. Introduction

Hidden Markov model (HMM) based text-to-speech (TTS) [1] provides a framework for flexible modeling and synthesis of different voices and voice qualities. Voice quality can be understood in many ways, but in this paper, we consider voice quality related to vocal fold vibration. Widely accepted types of phonation include modal, falsetto, breathy, and vocal fry [2]. In this paper, we focus on the case of vocal fry.

Vocal fry (glottal fry, creaky voice) refers to a voice quality characterized by a low and rather irregular rate of vocal fold vibration. Fundamental frequency (F0) is lower than in modal voice (normal phonation). A typical use of vocal fry is as an utterance final preboundary marker. Speech databases can inherently contain this kind of creaky endings. Using vocal fry at the end of an utterance in a proper way in TTS makes the result more natural.

For HMM-based TTS, speech must be parameterized. A widely used speech model is source-filter decomposition. Voice quality is mainly manifested in laryngeal features, i.e. in the source. A vocoder using a binary voiced/unvoiced decisions without any other measure of the source signal cannot capture the voice quality. In the HMM-based Speech Synthesis System (HTS) [3] version 2.1, mixed excitation [4] provided by STRAIGHT analysis and synthesis framework [5] is used. The use of mixed excitation provides a more natural speech quality compared to early versions using only binary voiced/unvoiced decision without any aperiodicity measure. Alternatively, good synthesis quality can be obtained by modeling the glottal source from the speech production point of view (e.g. [6]).

Overall, the success of the HMM-based speech synthesis depends on the success of the parameterization. If the speech analysis fails in extracting certain parameters, the synthesis cannot perform well either. This is also a problem with speech containing vocal fry. Due to differing phonation of modal voice and vocal fry, the F0 extraction algorithms finetuned for modal speech may fail in extracting F0 contours for the vocal fry segments. If erroneous F0 values, or no F0 values at all, are extracted for the training data, no proper models can be built. The averaging effect of HMM clustering even causes the problem to spread on longer segments in synthesis. The same problem applies for the mixed excitation measures.

A widely used analysis and synthesis framework is STRAIGHT [5]. It provides a high quality estimate of the speech spectrum envelope. In the current version of STRAIGHT, modeling of mixed excitation is also supported. According to our experiments, F0 and aperiodicity analysis of the standard STRAIGHT configuration tend to fail in vocal fry degrading the synthesis quality. Additionally, not all the voices are modal enough for high quality aperiodicity parameterization in STRAIGHT analysis. Thus, the usage of more non-parametric (waveform-based) features for the glottal source can be a better choice. We propose to replace the F0 and aperiodicity of STRAIGHT with a more waveform-based parameterization while still using the STRAIGHT spectrum. Aperiodicity is replaced with a model of two-band voicing in a similar manner to [7]; the frequency band is divided into a lower voiced frequency band and an upper unvoiced frequency band based on the estimated voicing cut-off frequency. In this paper, we propose the use of a voicing cut-off frequency estimation based on minimizing the resulting modeling error. A simple carefully set cut-off frequency value is shown even to improve the analysis/synthesis performance. This parameterization is able to provide consistent feature extraction also for the fry segments while maintaining the high-quality for other segments.

This paper is organized as follows. Section 2 describes the characteristics of vocal fry and the problems it causes in STRAIGHT analysis. The proposed waveform-based estimation for F0 and voicing cut-off frequency are described in Section 3. Section 4 presents the evaluation of the proposed parameterization in both analysis/synthesis and HMM-based TTS framework. Section 5 concludes the paper.

## 2. Vocal fry in speech analysis

### 2.1. Vocal fry characteristics

Vocal fry [8] is a voice quality characterized by a low and rather irregular rate of vocal fold vibration. Depending on the language, vocal fry may have a role as a functional, idiolectal, or emotional feature in speech. It can indicate e.g. different lex-

ical meanings, end of an utterance or a turn, pause filling, or speaker’s uncertainty or hesitation [9].

Fundamental frequency of vocal fry is below the fundamental frequency of a modal voice, typically below 70–90Hz for both female and male speakers [9]. Glottal excitation in vocal fry consists of a short open phase and a longer closed phase [2]. Speech waveform is characterized by strong damping between two adjacent glottal pulses [8]. Long pitch periods combined with strong decay between the excitation pulses make single pulses detectable. This results in a popping sound that is characteristic of vocal fry.

## 2.2. STRAIGHT analysis

In STRAIGHT analysis [5], speech waveform is decomposed into a F0 contour and a spectrum with no periodic interferences in time or frequency domain. In the current version [10], mixed excitation is also supported by introducing an aperiodicity map estimated from the residuals between harmonic components.

The spectrum is estimated from the speech signal by using pitch-adaptive time windows and complementary windows for reducing periodic interferences and phase variations of harmonic components in time domain. In addition, inverse filtering is used for removing frequency domain interference while preserving harmonic components [5]. The extraction algorithm for F0 and aperiodicity measure is described in [11]. In the first phase of this two-step extraction procedure, an instantaneous frequency-based estimate is produced. The selection of the best F0 estimate is done based on the maximum carrier-to-noise (C/N) ratio. In the second phase, the initial estimate is refined by harmonic component analysis. Furthermore, an estimate for the signal aperiodicity is produced.

## 2.3. STRAIGHT performance for vocal fry

According to [11], the F0 extraction performance of the STRAIGHT analysis in female speech is competitive with or even better than in the conventional methods. However, for the Finnish speech database used in the evaluations of Section 4, vocal fry endings, occurring in part of the utterances, cause problems in F0 extraction. In preliminary evaluation, generally no correct F0 value was extracted for fry segments. These segments were typically detected as unvoiced even when the F0 threshold was set to significantly low level. Based on manual extraction, F0 of 40–70Hz typically occurred in the fry segments. Most of the time, the algorithm was able to find suitable initial candidates for F0 in vocal fry segments. Nonetheless, these segments were typically classified as unvoiced. Problems in F0 extraction also affect the aperiodicity computation, further reducing the re-synthesis quality.

An example of failed F0 extraction in vocal fry is presented in Figure 1. Figure 1 represents an utterance final signal waveform for a sentence ending with the words ‘...tahi muualla’ (vocal fry part underlined in the transcription) and its F0 contour extracted by STRAIGHT. Even though all the phonemes of the last word are voiced, no F0 is extracted for the interval 5.95–6.15s. Manually extracted fundamental periods of this segment vary from 15ms to 26ms corresponding to F0 of 38–67Hz.

An improved extraction algorithm integrating estimates from instantaneous frequency-based extraction and interval-based extraction is introduced in [12]. Even when using this method, the vocal fry segments of the database were often detected either as unvoiced or voiced with too high F0 value.

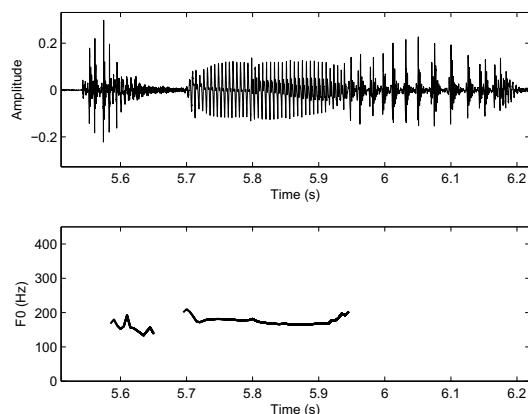


Figure 1: *Waveform and F0 contour extracted by STRAIGHT for an utterance final segment ‘...tahi muualla’.*

## 3. F0 and voicing estimation

In the proposed approach that is aimed for robust processing of vocal fry, the estimation of F0 and voicing for a given frame of speech begins with frequency-domain pitch estimation. A variable-length window based on the previous F0 value is applied for calculating Fourier transform, and a frequency-domain metric is computed for all the integer-length pitch periods using a sinusoidal speech model matching based approach that partially follows the ideas presented in [13]. Then, a time-domain metric measuring the similarity between successive pitch cycles is computed for a fixed number of pitch candidates that received the best frequency-domain scores. The actual pitch estimate is obtained using these two metrics together with a pitch tracking algorithm that considers a fixed number of potential pitch candidates for each analysis frame. As the final step, the obtained pitch estimate for a given frame is further refined using a sinusoidal speech model matching based technique to achieve better than one-sample accuracy that is necessary for perceptually accurate F0 modeling.

Once the final refined pitch value has been estimated, the next step is to estimate voicing. Again, the estimation is performed using the frequency domain representation generated by applying variable-length windowing and fast Fourier transform (FFT). The voicing information is derived for the residual spectrum through the analysis of voicing-specific spectral properties separately at each harmonic frequency. In particular, a voicing likelihood is computed for every harmonic by estimating the degree of harmonic structure based on the correlation between the spectral shape of each band and the corresponding spectral shape of the variable-length window. Finally, to simplify the representation to a single voicing parameter, a voicing cut-off frequency is determined by minimizing the resulting modeling error.

Figure 2 illustrates the extracted F0 contour for utterance final vocal fry extracted by the proposed method. The sentence is the same as in Figure 1. In Figure 2, most of the fry (5.95–6.15s) is classified as voiced.

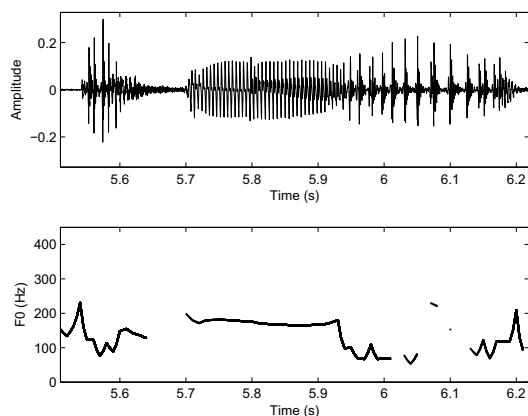


Figure 2: Waveform and F0 contour extracted by the proposed method for an utterance final segment ‘...tahi muulla’.

## 4. Evaluations

### 4.1. HMM-based speech synthesis system

Model training and parameter generation were done using HTS [3] version 2.1. The adaptation of HTS for Finnish is described in more detail in [14]. A Finnish speech database of 80 minutes recorded by a female speaker was used for the HMM training. The database sentences were recorded in a vivid style leading to the occurrence of the vocal fry at the end of the utterances. Features employed in the training are described in Section 4.2. All the features were augmented with their deltas and delta-deltas.

### 4.2. Speech parameterization

For both the analysis/synthesis and HMM-based TTS tests, two different parameterizations were considered:

- Baseline: STRAIGHT spectrum represented as Melcepstral coefficients of order 39, STRAIGHT F0, and mean STRAIGHT aperiodicity of the five bands as in [4].
- Proposed: STRAIGHT spectrum represented as Melcepstral coefficients of order 39, proposed F0, and proposed two-band voicing cut-off frequency.

In both cases, the spectrum was extracted using the F0 estimation provided by STRAIGHT. An analysis update interval of 5ms was used for both approaches.

### 4.3. Quality evaluation

Eight naive listeners rated the speech quality in a Mean Opinion Score (MOS) test. Recorded speech (Natural) as well as the result of the analysis/synthesis (A/S) for the baseline (A/S-baseline) and proposed parameterization (A/S-proposed) were evaluated with 10 same sentences for each. In addition, 14 sentences from HMM-based speech synthesis using the baseline (HTS-baseline) and proposed (HTS-proposed) parameterization were rated. Both sentence sets were selected randomly and the selection was not restricted by the requirement of vocal fry occurrence. Because durations play a significant role in Finnish speech, the same phoneme durations were assigned for both sentences in HMM-based TTS by using the same duration tree in param-

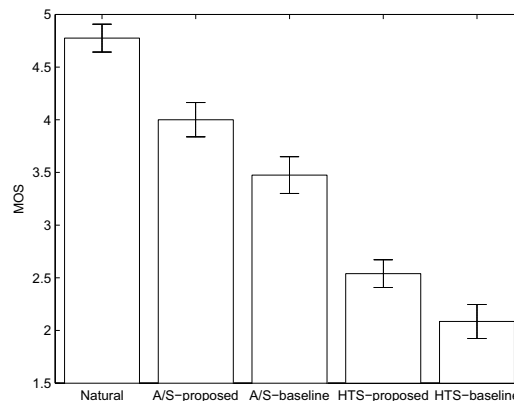


Figure 3: Mean opinion scores with 95% confidence intervals for subjective evaluation using rating 1-5 (1=Bad, 2=Poor, 3=Fair, 4=Good, 5=Excellent).

ter generation. The readers can listen to the audio samples available at [http://www.cs.tut.fi/sgn/arg/interspeech09/vocal\\_fry.html](http://www.cs.tut.fi/sgn/arg/interspeech09/vocal_fry.html).

All sentences (62 in total) were evaluated together and presented to the subjects in a random order. The mean scores for the MOS test are shown in Figure 3 with 95% confidence intervals. The absolute values are not important but rather the distances and confidence intervals between the approaches.

In both A/S and HMM-based speech synthesis, the proposed parameterization can be concluded to result in better quality when considering 95% confidence intervals. The results confirm that replacing the glottal source parameterization does not degrade the quality but actually improves it. In A/S samples, the vocal fry plays only a relatively minor role and in our opinion, the improvement shows that the proposed approach provides a more suitable aperiodicity measure also for non-fry segments.

### 4.4. Comparison test

The relative quality of the proposed and baseline parameterization in HMM-based TTS was further evaluated using a pairwise comparison test. In the evaluation, eight listeners were asked to compare 16 randomly ordered synthesized sample pairs. Each pair consisted of a sentence generated using both the proposed and the baseline parameterization. The evaluation sentences were selected randomly as in Section 4.3. The use of global variance mainly improves the speech spectrum [15] and may cause problems in F0. Global variance was not therefore considered in the parameter generation. Instead, postfiltering was used.

The results of the comparison test are shown in Table 1. As can be seen, the proposed approach clearly outperformed the baseline system. The proposed parameterization was preferred 89.8% of the time while the preference percentage for the baseline approach was 2.3%. The main reason for the poor performance of the baseline was the major problems in vocal fry parts. The effect of the unvoiced classification of the fry parts was spread over the whole last word.

An example of the HMM-based synthesis result for both (a) the baseline speech model and (b) the speech model using the proposed excitation parameterization is illustrated in Figure 4.

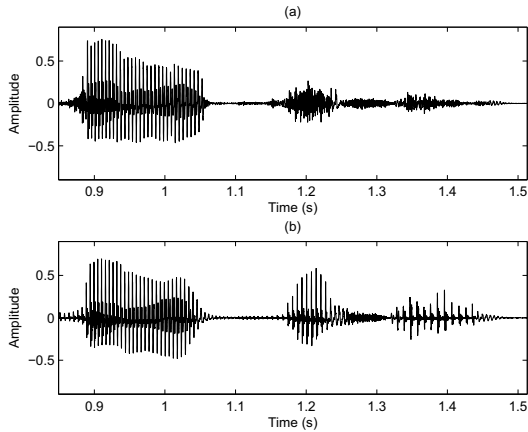


Figure 4: Utterance final vocal fry in a synthesized sentence ‘...ollut asiaa’ using (a) the baseline (b) the proposed parameterization.

The synthesized sentence was not included in the training data. In (b) the vocal fry of the last word has been successfully synthesized whereas in (a) the whole last word is rather noisy and lacks periodicity although the word consists mainly of vowels. In modal phonation, no major differences exist.

Table 1: Comparison test preference and 95% confidence intervals for the proposed and baseline parameterizations in HMM-based speech synthesis.

	Percentage
Preference for proposed	89.8% $\pm$ 5.3%
No preference	7.8% $\pm$ 4.7%
Preference for baseline	2.3% $\pm$ 2.7%

## 5. Conclusions

In speech synthesis, it is beneficial to be able to use different kinds of speech databases with expressive quality. Conversational or vivid databases usually contain non-modal speech like vocal fry that makes the parameterization more challenging, and cause problems for conventional approaches such as STRAIGHT. Due to the averaging effect of model training and clustering, the problems become even more emphasized in HMM-based speech synthesis. In this paper, we have proposed a new approach for F0 and aperiodicity estimation to be combined with the STRAIGHT spectrum in HMM-based TTS. The proposed method is especially suitable for robust processing of vocal fry. It was verified through listening tests that the new parameters did not cause quality degradation in analysis/synthesis. Further, the proposed configuration was shown to clearly improve the quality of HMM-based speech synthesis over the baseline configuration.

## 6. Acknowledgements

This work was supported by the Academy of Finland, (application number 129657, Finnish Programme for Centres of Excellence in Research 2006- 2011).

## 7. References

- [1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in *Eurospeech*, 1999, pp. 2347–2350.
- [2] D. Childers and C. Lee, “Vocal quality factors: Analysis, synthesis, and perception,” *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, 1991.
- [3] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, “The HMM-based speech synthesis system (HTS) version 2.0,” in *6th ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.
- [4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Mixed excitation for HMM-based speech synthesis,” in *Eurospeech*, 2001, pp. 2263–2266.
- [5] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, pp. 187–207, 1999.
- [6] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, “HMM-based Finnish text-to-speech system utilizing glottal inverse filtering,” in *Interspeech*, 2008, pp. 1881–1884.
- [7] S.-J. Kim, J.-J. Kim, and M. Hahn, “HMM-based Korean speech synthesis system for hand-held devices,” *IEEE Transactions on Consumer Electronics*, vol. 52, no. 4, 2006.
- [8] H. Hollien and R. Wendahl, “Perceptual study of vocal fry,” *J. Acoust. Soc. Am.*, vol. 43, no. 3, pp. 506–509, 1968.
- [9] A. Iivonen, “Creaky voice as a prosodic feature in Finnish,” in *Nordic Prosody, IX Conference*, 2004, pp. 137–146.
- [10] H. Kawahara, “STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds,” *Acoust. Sci. & Tech.*, vol. 27, no. 6, pp. 349–353, 2006.
- [11] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity,” 1999, pp. 2781–2784.
- [12] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Irino, “Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT,” in *Interspeech*, 2005, pp. 537–540.
- [13] R. McAulay and T. Quatieri, “Pitch estimation and voicing detection based on a sinusoidal speech model,” in *ICASSP*, 1990, pp. 249–252.
- [14] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, “Evaluation of Finnish unit selection and HMM-based speech synthesis,” in *Interspeech*, 2008, pp. 1853–1856.
- [15] T. Toda and K. Tokuda, “Speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” in *Interspeech*, 2005, pp. 2801–2804.