INTERSPEECH 2009 BRIGHTON

Nearly Perfect Detection of Continuous F_0 Contour and Frame Classification for TTS Synthesis

Thomas Ewender, Sarah Hoffmann and Beat Pfister

Speech Processing Group
Computer Engineering and Networks Laboratory
ETH Zurich, Switzerland
{ewender,hoffmann,pfister}@tik.ee.ethz.ch

Abstract

We present a new method for the estimation of a continuous fundamental frequency (F_0) contour. The algorithm implements a global optimization and yields virtually error-free F_0 contours for high quality speech signals. Such F_0 contours are subsequently used to extract a continuous fundamental wave. Some local properties of this wave, together with a number of other speech features allow to classify the frames of a speech signal into five classes: voiced, unvoiced, mixed, irregularly glottalized and silence.

The presented F_0 detection and frame classification can be applied to F_0 modeling and prosodic modification of speech segments in high-quality concatenative speech synthesis.

Index Terms: continuous F_0 contour estimation, globally optimized F_0 contour, voiced-unvoiced classification, irregular glottalization, vocal fry

1. Introduction

Most high-quality speech synthesis systems are based on concatenation of natural speech segments. The quality of these systems can benefit significantly from a good estimation of the fundamental frequency contour of speech signals. It allows to optimize on one hand the prediction of the target prosody of the speech to be synthesized and on the other hand the selection and prosodic modification of the segments to be concatenated.

In terms of prediction of target prosody in general and F_0 contours in particular, a most successful approach is based on a recurrent multi-layer perceptron as described in [1] or in [2]. There it was also shown, that much better models could be achieved with continuous F_0 contours rather than with only piecewise defined ones. Hence, we need a possibility to estimate for a given speech signal an F_0 contour that is accurate in voiced parts and reasonably smooth in unvoiced parts.

When segments are concatenated they have to undergo some prosodic modifications, no matter if the size of the speech corpus is minimal as in diphone synthesis or large as in unit selection synthesis. Such modifications can for example be performed by means of time or frequency domain PSOLA (see e.g. [3]). A prerequisite is again an accurate estimate of the F_0 contour. But prosodic modification also depends on signal properties such as voiced, unvoiced, mixed, irregularly glottalized (typically creaky voice segments) or silent. Therefore, each frame of the speech signal has to be assigned to one of these classes.

In the following section, we will show a new, robust but nevertheless very accurate method for the determination of a continuous F_0 contour for good quality speech signals as they are used in concatenative TTS synthesis. In section 3 an ANN-based classifier will be presented that assigns each speech frame to one of the five above mentioned classes.

2. Estimation of the T_0 contour from a high-resolution cepstrogram

Our approach to estimate a continuous F_0 contour is motivated by the fact that humans can easily and reliably "see" the T_0 contour of a speech signal from a suitably drawn cepstrogram such as the one shown in Figure 3: The contour has to follow the strong cepstral peaks (i.e. the bright tracks), has to be somewhat smooth and should not make unreasonable detours in completely unvoiced regions.

This task can be considered as an optimization problem, namely to find the optimal curve along the cepstral peaks while the curve has to meet some constraints at the same time. As constraints we use the probability distribution of the local declination and curvature that has been estimated from natural F_0 contours, as shown in Section 2.2. The optimization is detailed in Section 2.3.

2.1. Computation of high-resolution cepstrogram

In order to get the high-resolution cepstrogram, we first compute the logarithmic power density for every frame:

$$S(k) = log(\frac{1}{NU}|X(k)|^2)\,, \qquad 0 \le k \le N-1 \qquad (1)$$

where $X(k)$ is the discrete Fourier transform of the frame, N the window length and U a constant to compensate for the window function. The power density is defined for the discrete frequencies $f_k = f_s k/N$, with $k = 0, 1, \ldots, N-1$.

Subsequently we eliminate the frequency components of the spectrogram that are higher than $f_b = f_s b/N$. In order to get a cepstrogram with sufficiently high resolution we apply padding to the spectrogram

$$\begin{aligned} S'(k) &= S(k)\,, & 0 \le k \le b \\ S'(k) &= S(b)\,, & b < k < M-b-1 \\ S'(k) &= S(M-k-1)\,, & M-b-1 \le k < M \end{aligned} \qquad (2)$$

where M is the number of points used for the inverse Fourier transform to calculate the cepstrum and $N/(Mf_s)$ is the quefrency resolution of the cepstrogram. For the cepstrogram shown in Figure 3, we used a 50 ms Hamming window at a sampling frequency of 22.05kHz, f_b and M were set to 5 kHz and 8192, respectively.

2.2. Estimating the probability distribution of local declination and curvature

We define the local declination $d(t)$ and the curvature $c(t)$ of the discrete time sequence $q(t)$ with sampling points at every time interval T_s as follows:

$$d(t) = \frac{(q(t) - q(t-2T_s))/2 + q(t) - q(t-T_s)}{2T_s} \quad (3)$$

$$c(t) = \frac{q(t) - 2q(t-T_s) + q(t-2T_s)}{T_s} \quad (4)$$

In our case $q(t)$ stands for a sequence of logarithmic quefrencies that has been evaluated from a speech signal with a frame shift of T_s. Hence, d and c are relative changes and will be expressed in the following as %/s (percent per second).

The 2-dimensional probability distribution of the declination and curvature has been estimated from the voiced parts of some 17 hours of speech from various speakers and languages. In a first step we have detected F_0 of all signals by means of an algorithm similar to YIN (see [4]) and converted the F_0 values to logarithmic T_0 values. Then for triples of consecutive log T_0 values the declination and curvature has been evaluated. An overview of the resulting pairs of declination d and curvature c is shown as a normalized histogram in Figure 1.

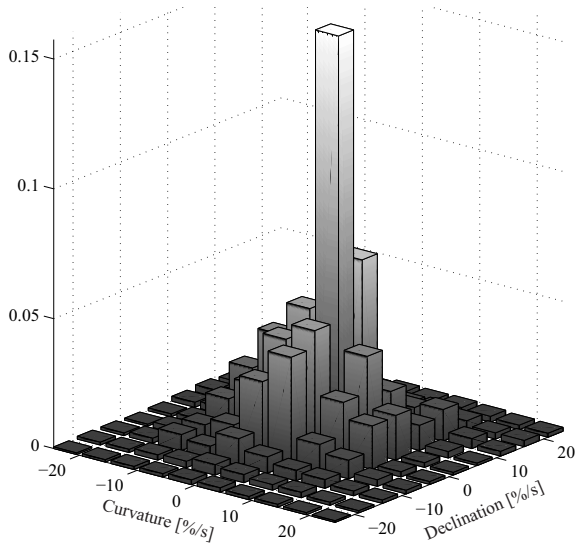

Figure 1: *Empirical probability distribution of the local declination and curvature of natural T_0 contours*

Since an empirical probability distribution as shown in Figure 1 is not practically usable, we have approximated it with a 2-dimensional Gaussian mixture model with two mixture components. The probability density function $p(d, c)$ of the resulting model is shown in Figure 2.

2.3. Finding the most probable continuous T_0 contour

To determine the optimal T_0 contour in the log cepstrogram $C(t, l)$, where t and l are the discrete time and log quefrency, respectively, we devised a Viterbi-like procedure (inspired by [5] and [6]). This procedure evaluates the globally optimal sequence of log quefrency values over all discrete times t. The optimization is based on a local score $\alpha(t, l)$ that considers $C(t, l)$

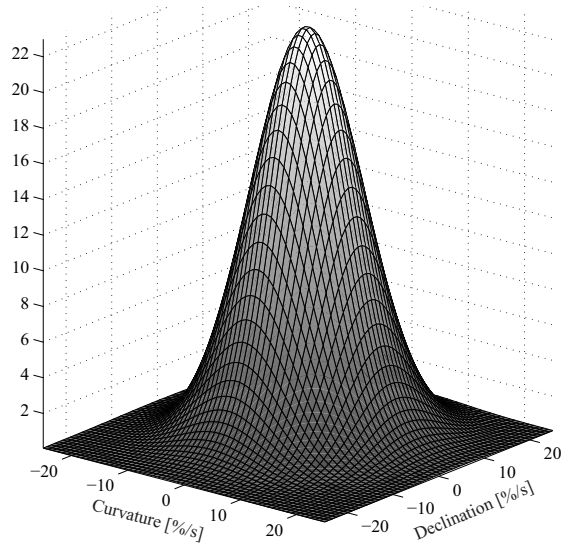

Figure 2: *Probability density function $p(d, c)$ of the GMM approximating the 2-dimensional distribution of local declination d and curvature c of natural T_0 contours*

and the likelihood of the local declination and curvature. More formally, the local score is defined as

$$\alpha(t, l) = p(d, c) \cdot e^{wC(t,l)} \quad (5)$$

whereby w is a weighting factor that equals the signal power of the corresponding frame. Note that $p(d, c)$ depends on the two preceding points of the T_0 contour (cf. equations 3 and 4). The overall score of the optimal T_0 contour can then be found with the following iteration over all discrete quefrency and time values:

$$\delta(t, l) = \max_k \{\delta(t-T_s, k) \cdot \alpha(t, l)\}, \quad (6)$$

where $\delta(t, l) = 1$ for $t \leq 0$. To find the most probable sequence of (t, l) pairs at the end of the recursion it is necessary to store during the recursion the optimal predecessor l for every time t and quefrency l in $\Psi(t, l)$. Like in the Viterbi algorithm, the optimal sequence of (t, l) pairs can then be found by starting at the end point and going iteratively backwards. A resulting T_0 contour is shown as a continuous line in the cepstrogram in Figure 3.

The computation of this search algorithm is done in log domain for numerical reasons.

3. ANN-based frame classification

It was outlined in the introduction, that besides an accurate F_0 contour we need also information about the frame properties such as voicing, pitch regularity, etc. We have therefore defined five classes of frames that must be treated differently in prosodic modifications. These five classes are:

voiced: Speech with clearly perceptible voicing; harmonic signal; not noisy; low frequencies dominant (typical phonemes: vowels, nasals)

unvoiced: Speech without perceptible voicing; noisy; high frequencies above 2 kHz are dominant (typical phonemes: fricatives, unvoiced plosives)

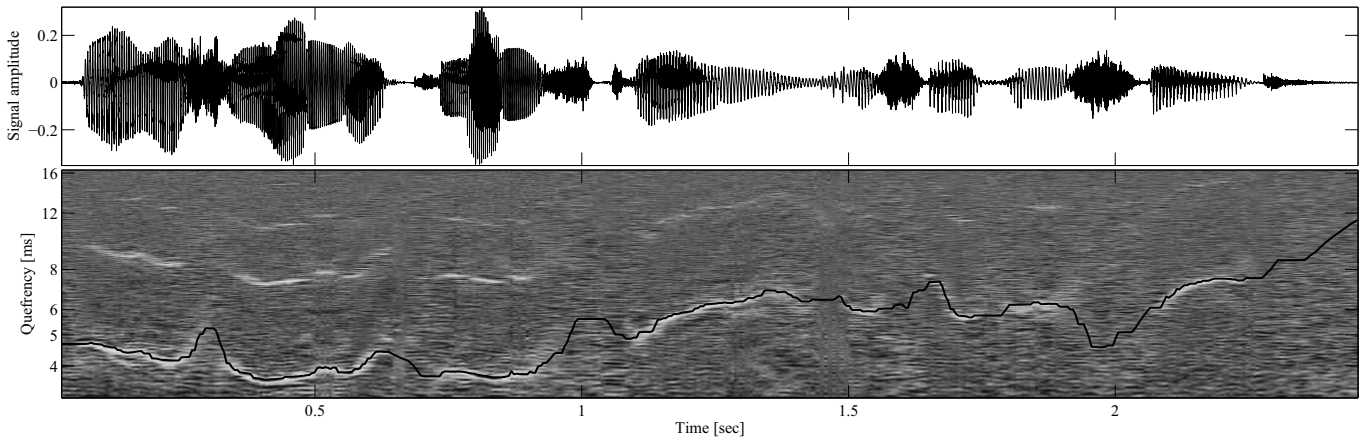

Figure 3: *Speech signal and the corresponding high-resolution cepstrogram with detected T_0 contour, drawn on a logarithmic quefrency axis. Because $F_0 = 1/T_0$, the F_0 contour can easily be achieved by vertically flipping the shown T_0 contour.*

mixed: Speech with voicing and noise; only lower harmonics visible in spectrum; higher spectral components noisy (typical phonemes: voiced fricatives; frequently in voiced-to-unvoiced transitions)

irregular: Speech with irregularly spaced glottal pulses; no significant fricative components; low frequencies dominant; also known as creaky or stiff voice or vocal fry; very frequent in some voices or languages (occurs often in voiced plosives and towards the end of utterances, when F_0 drops to very low frequencies or e.g. in Mandarin low tones)

silence: Signal segments with low energy; virtually not audible

The above classification criteria are primarily motivated by the application (i.e. to perform prosodic modifications) rather than by linguistic arguments.

Such a classification can be realized with a feed-forward ANN. Important input information for the classifier can be derived from the local properties of the fundamental waveform, as shown in Section 3.2. First we will explain, how we generate a virtually continuous fundamental wave for a whole speech signal.

3.1. Generating the fundamental wave

The fundamental wave can be achieved from the convolution of the speech signal with a Hamming window of the size T_0. This is a low-pass filter with a cut-off frequency of about F_0. Since T_0 varies along the speech signal, the size of the Hamming window has to be adapted continuously, i.e., not only for each frame, but for each sample. Because a T_0 contour, as it results from the optimization described in Section 2.3 is specified with one value per frame, it has to be interpolated to get a T_0 value for each sample of the speech signal.

The resulting fundamental wave is completely smooth and particularly shows no discontinuities at frame boundaries. A segment of such a fundamental wave is shown in Figure 4.

3.2. Properties of the fundamental waveform

It can easily be seen that the above sketched convolution produces a signal that is indeed equal to the fundamental wave of a quasi-stationary harmonic signal like voiced speech. This fundamental wave is close to sinusoidal and its period changes only slowly. Conversely, in clearly unvoiced sections of the speech signal the fundamental wave gets very irregular, in terms of amplitude as well as with respect to the period. Also sections of vocal fry are clearly visible, because their fundamental wave typically is neither close to sinusoidal nor regular periodic.

The local properties of the fundamental waveform can be described by a number of simple features that will be used by the frame classifier. These and further features will be sketched in Section 3.3.

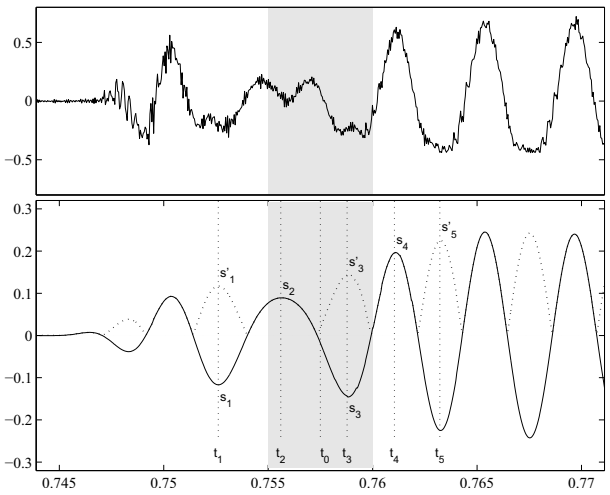

Figure 4: *Speech segment and derived fundamental wave; the current frame is marked in grey; its center is at time t_0*

3.3. Classification features

The features used for the classifier are outlined below. Features number 6 to 10 describe the degree of regularity of the fundamental wave. The times and points indicated refer to the segment shown in Figure 4. The described features are meant for the frame interval 0.755 to 0.76 s and t_0 designates the middle of this frame.

1. Zero crossing rate
2. Speech signal power (in logarithmic scale)

3. Spectral tilt (first mel frequency cepstral coefficient)
4. Dominance of central frequencies (second mel frequency cepstral coefficient)
5. Periodicity: value of the cepstrogram at quefrency P_0
6. Amplitude of the fundamental wave in terms of distance between the two line segments defined by the points s_2 and s_4 and by the points s_1 and s_3 at time t_0
7. Dynamics of the fundamental waveform: $|1-f_1|$, where $f_1 = (t_2-t_1)/(t_3-t_2) \cdot T_0(t_3)/T_0(t_1)$ and $T_0(t)$ is the period of the speech signal at time t, estimated as described in Section 2.3
8. Similar to feature 7, but using time points t_2, t_3 and t_4: $|1-f_2|$, where $f_2 = (t_3-t_2)/(t_4-t_3) \cdot T_0(t_4)/T_0(t_2)$
9. Regularity of the fundamental waveform: $|1-f_3|$, where $f_3 = 4(t_3-t_2)/(T_0(t_3) + T_0(t_2))$
10. Irregularity of increase or decrease of fundamental wave amplitude: mean square error of the quadratic regression of the points s'_1, s_2, s'_3, s_4 and s'_5

3.4. Training of the ANN-based classifier

For the development of the ANN-based classifier we used 6 minutes of studio quality speech signal from 20 different voices covering a range of 12 European and Asian languages. These speech signals contain a great diversity of voice qualities including plenty of creaky and stiff voice segments.

We manually classified these speech signals into segments of the five classes given at the beginning of Section 3 by looking at the waveform, spectrogram, the fundamental wave, the pitch contour and by listening to the speech segments.

The manually classified speech signals have been split into a training and a test set, whereby no speaker was in both sets. This allows to estimate the speaker-independent classification rate.

Finally, we trained a fully connected 2-layer ANN with 10 inputs, 6 nodes in the hidden layer and 5 nodes in the output layer. For the training the back-propagation was used with randomly selected sub-epochs. We used no evaluation set, as it is generally used for stopping the training at the optimal point, because previous experiments have shown that the training of our rather small ANN is not critical in terms of over-fitting.

4. Evaluation

We refrained from a commonly done comparison of our F_0 results with laryngograph-based estimates, because unclear frames are usually excluded for objectivity reasons from such tests. However, unclear frames are most interesting with respect to our application (prosodic modification of speech). Furthermore, it is questionable whether for such frames objectively more reliable values can be estimated from a laryngograph signal at all. Instead of a not very helpful comparison we demonstrate the quality of our results by means of examples on the web.

Table 1: *Classification accuracy in %*

	unvoiced	silence	voiced	mixed	irregular
unvoiced	85.28	2.60	0.00	9.68	2.44
silence	1.81	93.15	0.04	1.07	3.93
voiced	0.01	0.03	90.69	3.38	5.89
mixed	8.56	1.14	4.12	76.60	9.58
irregular	1.88	2.60	3.86	7.62	84.04

As for the frame classification we evaluated the accuracy of the ANN in a strictly framewise comparison of the ANN output with the manual frame classification. The mean relative classification rate was 90.49%. The results per class can be seen in Table 1. By allowing a shift tolerance of 5ms (one frame) for phone transitions, the total accuracy increased 1.95 %.

5. Discussion

The presented F_0 detection is based on a clear mathematical model and on statistical properties acquired from a large speech corpus. The global optimization (in contrast to piecewise as in [5] and [6]) gives more robust results and also does not need any post-processing of the resulting F_0 or T_0 contours.

As can be seen from table 1, the voiced/unvoiced decision is nearly done perfectly. If confusions occur, they do between those classes that share signal qualities, as unvoiced and mixed and voiced and irregular. Inspection showed that these confusions mostly concern border cases where arguments for both classes are present.

6. Conclusions

The estimated F_0 contours are virtually perfect: we have not seen any errors in all manually inspected F_0 contours. Furthermore, the F_0 detection works also well for expressive speech with very high F_0 dynamic which several authors reported to be a problem for their algorithms.

Our F_0 contours are very well suited for the type of F_0 modeling described in [2] and for the generation of the fundamental wave of a speech signal in Section 3.1. We are convinced that based on the achievements reported in this paper, prosodic modification of speech signals will be possible without any audible artifacts.

7. Acknowledgments

This work was supported by the Swiss Innovation Promotion Agency CTI. We cordially thank Cédric Schaller who contributed to this work with his masterthesis.

8. References

[1] C. Traber, "Svox: The implementation of a text-to-speech system for german," Ph.D. dissertation, No. 11064, Computer Engineering and Networks Laboratory, ETH Zurich, TIK-Schriftenreihe Nr. 7 (ISBN 3 7281 2239 4), March 1995.

[2] H. Romsdorfer, "Polyglot text-to-speech synthesis: Text analysis & prosody control," Ph.D. dissertation, No. 18210, Shaker Verlag Aachen (ISBN 978-3-8322-8090-1), February 2009.

[3] F. Charpentier and E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," in *Proceedings of Eurospeech'89*, 1989, pp. 13–19.

[4] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[5] H. Kawahara and A. de Cheveigné, et al., "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Proceedings of Interspeech*. ISCA, 2005, pp. 537–540.

[6] D. Joho, M. Bennewitz, and S. Behnke, "Pitch estimation using models of voiced speech on three levels," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Honolulu, Hawaii, USA, April 2007, pp. 1077–1080.