

HIGH-QUALITY PROSODIC MODIFICATION OF SPEECH SIGNALS

Beat Pfister

Computer Engineering and Networks Laboratory (TIK)
Speech Processing Group, ETH, CH-8092 Zurich, Switzerland

ABSTRACT

The aim of this work was to develop a procedure that allows prosodic modifications of speech signals without impairing the quality. The developed procedure is based on the Fourier analysis/synthesis technique with several improvements on the analysis side, such as the analysis of signals with rapidly changing F_0 and the analysis of weak spectral components. Also for the modification of the short-time spectrum and for the reconstruction of the speech signal some new methods have been introduced. The most important one, in terms of speech quality, is the way of phase compensation that limits the absolute time shift to half the pitch period.

The developed procedure is used in our high-quality text-to-speech synthesis system that is based on concatenation of prosodically modified diphones.

1. INTRODUCTION

Text-to-speech (TTS) synthesis based on concatenation of natural speech segments, e.g. diphones, requires an algorithm for prosodic modification of these segments. Linear prediction (LP cf. [1]) provides an easy way to independently control the prosodic parameters duration, fundamental frequency (F_0), and intensity of the speech segments. Unfortunately, LP causes a clearly noticeable quality loss of the synthesized speech.

The well-known LP replacement used in today's TTS systems is PSOLA, a method based on pitch-synchronous overlap-add of signal periods that are modified either in the frequency or time domain (see e.g. [2] and [4], resp.). For very accurately pitch-segmented speech signals, the PSOLA method exceeds the LP quality by far. The speech quality, however, is very sensitive to deficiencies of the pitch segmentation. It is therefore advantageous to use a fixed frame rate procedure that needs no pitch segmentation at all.

An obvious solution is to use the short-time Fourier analysis/synthesis, as proposed e.g. in [3]. High-quality modification can only be achieved, however, if the following requirements are met: First of all, we need an accurate estimate

of the true short-time spectrum $X_{m,i} = \{a_{m,i}, f_{m,i}, p_{m,i}\}$, where m and i are the frame and the spectral component indices, resp. Additionally, appropriate methods for modifying the spectrum and for reconstructing the signal must be available. In the subsequent sections we will present our solutions.

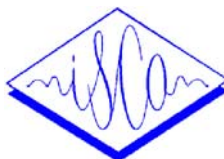
2. ESTIMATING THE TRUE SPECTRUM

The basic idea of the Fourier analysis/synthesis approach to the prosodic modification of speech signals, in particular with respect to F_0 and duration, is, to change the F_0 by scaling the frequency axis and the duration by summing up sine wave components of the desired length. This is not feasible with a standard short-time Fourier transform, because the result of the Fourier transform of a windowed signal is equal to the convolution of the signal spectrum with the spectrum of the window. The question is therefore: what is the spectrum that can be used for prosodic modifications (we term this the *true spectrum*), and how can it be evaluated.

It is well-known that a signal with period T_0 has a line spectrum with a line spacing of $F_0 = 1/T_0$. Furthermore, it has been shown, that the Fourier transform of a windowed periodic signal, where the Hamming window is at least $2.5T_0$ long, generally exhibits a relative maximum of the amplitude for every harmonic component. Taking these maxima as components of the true spectrum is not precise enough as shown in Figure 1. There are several obvious inaccuracies: the spectrum is not harmonic although the signal is periodic, weak components are not reliable, and there are many spurious peaks. These problems are associated with voiced speech, whereas unvoiced speech has turned out to be unproblematic. Therefore, the following considerations are related to voiced speech only.

2.1. Detection of weak spectral components

The signal reconstructed from the short-time spectrum of Figure 1 sounds exactly the same as the original one, unless



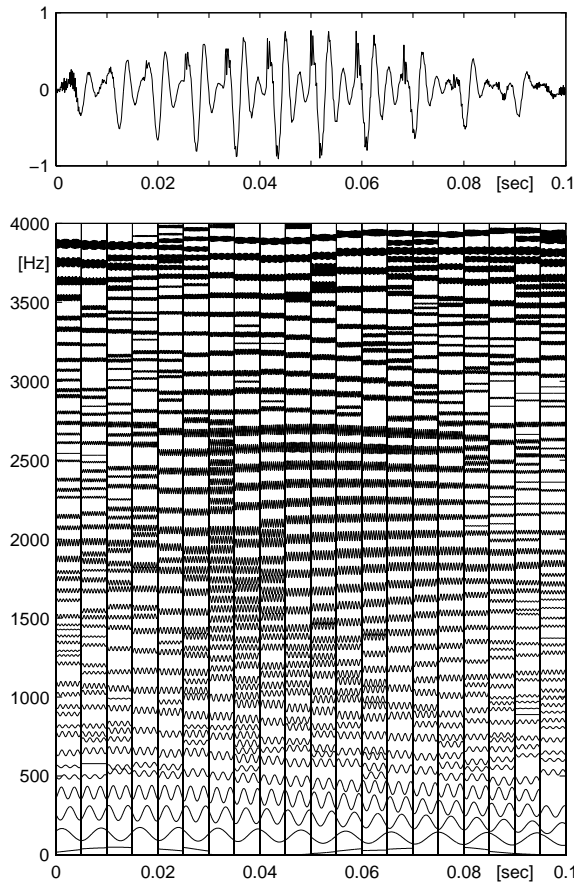


Figure 1: A speech signal with decreasing fundamental frequency ($\dot{F}_0 \approx -7.5$ oct/sec) and the corresponding short-time spectrum. Each relative maximum has been plotted as a short piece of a sine wave in the time frequency plane, in order to show time, frequency, amplitude (log), and phase information.

prosodic modifications are introduced. In particular, changing the F_0 may shift inaccurate, i.e. non-harmonic components into formants with a strong amplification. Simultaneously, reliably estimated strong components are shifted out of formant regions and consequently are attenuated. As a consequence, the frequency-scaled signal sounds hoarsely and not clearly articulated.

The analysis of artificial, stationary speech signals (signal generated from the spectral peaks of a single frame, see Figure 2a) using a high frame rate of 400 Hz demonstrated that small spectral peaks are somewhat noisy as shown in Figure 2b. In order to get a more reliable spectrum of a particular speech frame, we analyze a sequence of $2K + 1$ frames, whereby the center frame is identical to the frame in consideration and the frame shift T_f is very small. The peaks of the true spectrum are then evaluated by some sort of averaging the frequencies, the amplitudes and the phases (the phases must be unwrapped; cf. [6]). We have found that for $K = 3$ and $T_f = 1$ msec small spectral peaks can be estimated reliably enough as shown in Figure 2c.

Spurious peaks generally reside in the valleys between effective peaks. In our algorithm, peaks are searched iteratively, i.e. given a peak at f_{p_n} , then the highest peak in the interval $[f_{p_n} + 0.5 F_0 \dots f_{p_n} + 1.5 F_0]$ becomes peak p_{n+1} . Spurious peaks are thus automatically eliminated.

2.2. Analysis of signals with rapidly changing fundamental frequency

A second problem arises from the fast changing F_0 . Experiments with analytical, harmonic signals (i.e. with given spectrum) have shown that for frequency components higher than about $F_L = 2W_H F_0 / \dot{F}_0$, where W_H is the size of the analysis Hamming window in msec, F_0 the fundamental frequency in Hz, and \dot{F}_0 the variation of F_0 in oct/sec. This problem can be solved in either of two ways:

- Since \dot{F}_0 of the current frame can be estimated from the F_0 of the adjacent frames, we know that the signal frame is composed of sine waves with the same relative frequency change. Instead of the standard Fourier transform for stationary signals, the following transform can be used

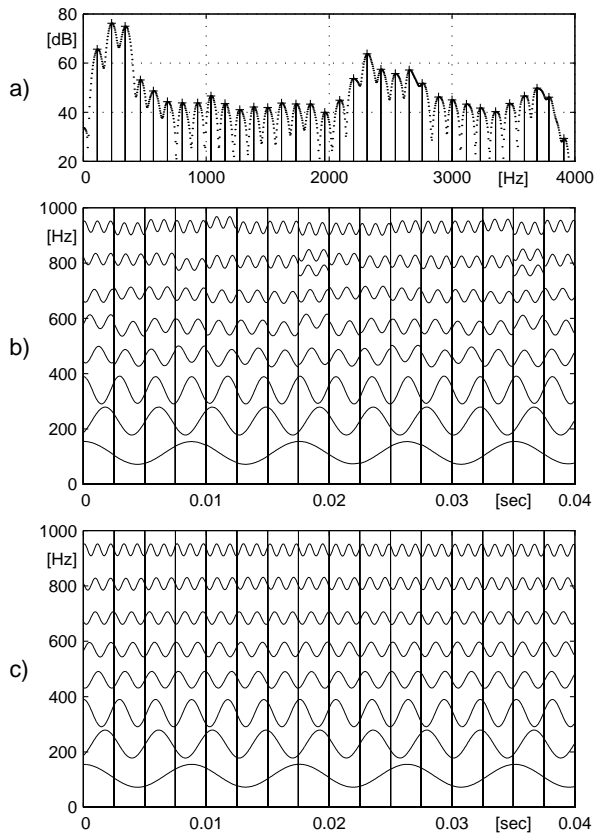


Figure 2: From the spectrum (top) the Fourier synthesis generates an exactly stationary signal. The standard short-time Fourier analysis of this signal exhibits unreliable results for weak components (middle). The results from the multi-window analysis are much better (bottom).

$$X(k) = \sum_{n=0}^{N-1} s(n) e^{-j2\pi k(n+\varphi(n))/N}, \quad (1)$$

where $\varphi(n)$ accounts for the relative frequency increase. The drawbacks of this transform are the lack of an efficient algorithm like the FFT, and inevitable aliasing effects near $F_s/2$.

- A signal frame with known \dot{F}_0 can be transformed into a frame with (approximately) constant F_0 by means of a sampling rate converter as shown in [6]. From this signal the harmonic components are estimated using the standard Fourier transform.

The spectral analysis with the improvements described in sections 2.1 and 2.2b) is shown in Figure 3.

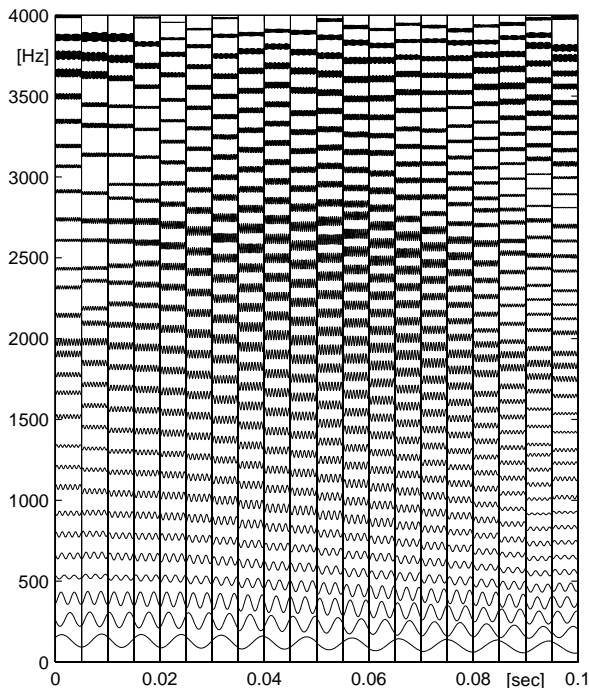


Figure 3: The improved spectrum analysis of the signal of Figure 1 shows the same relative frequency change of all harmonic components. In regions where frication or breathing noise is dominant, the spectral peaks are not regularly spaced, of course.

3. RECONSTRUCTING THE SPEECH SIGNAL

From the true short-time spectrum, the speech signal can be reconstructed either by using a set of parallel, controlled sine wave generators (like in [5]) or by applying an overlap-add technique (OLA). We consider the approach in [5] for two reasons as problematic: First, it is often not decidable

which of the frequency components of frame m and $m+1$ belong to the same sine wave generator, and second, scaling time and/or frequency in the time-phase plane, in order to make prosodic modifications of the signal, produces for not exactly harmonic frequency components a perceptually bad result. The reason is mainly the time shift, that can turn originally similar waveforms of neighboring frames into very different ones, as illustrated in Figure 4.

Furthermore, the time and pitch scaling scheme should also allow prosodic modification of discontinuous signals, e.g. concatenated diphones as used in a TTS system. That is why we decided to use the OLA method to reconstruct the signal. The perceptually optimal overlap size of the weighting windows is about 2–3 msec.

4. PROSODIC MODIFICATIONS

This section demonstrates how prosodic modifications of a speech signal that is given as short-time spectrum can be achieved. The time and pitch (frequency) modifications are controlled by the factors Z_T and Z_F , resp.

4.1. Time Scaling

Scaling the time axis while keeping the other signal properties constant means that the duration D of frame m becomes $D Z_T(m)$. In order to maintain the periodicity across the frame boundaries, the propagated time shift from frame m into frame $m+1$, i.e. $\Delta_T(m+1) = \Delta_T(m) + D \cdot (Z_T(m) - 1)$, has to be taken into account (a time shift of Δ_T means for each component a phase offset of $\Delta_{\varphi_i} = 2\pi f_i \Delta_T$). Depending on the values of $Z_T(m)$, the accumulated time shift may become large and cause the problem illustrated in Figure 4.

This problem can obviously be omitted as follows: Instead of shifting the signal according to the accumulated time shift, we shift the signal at most half of a pitch period to maintain the periodicity. In order to account for the fact that either

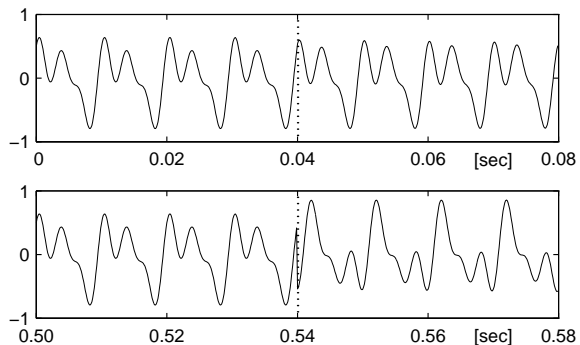


Figure 4: The upper part shows the signal given by the short-time spectrum $X_{m,i}$ with the frequency components $f_{1,i} = [100, 200, 300]$ and $f_{2,i} = [100, 201, 301]$ for a frame size of 40 msec. The lower part shows the discontinuity resulting from a time shift of 0.5 sec.

$F_0(m)$ and $Z_T(m)$ are constant across a frame, but variable over the whole signal, the phase compensation for the spectral components is given by the recursion in section 4.2 (with $Z_F(m) = 1$).

4.2. Pitch Scaling

Frequency and time scaling require similar phase compensations that can be combined in the following recursion:

$$\begin{aligned} \check{f}_{m,i} &= f_{m,i} Z_F(m) \\ \check{F}_0(m) &= F_0(m) Z_F(m) \\ \Delta_T(m) &= \Delta_{\varphi_o}^*(m) / (2\pi \check{F}_0(m)) \end{aligned} \quad (2)$$

$$\check{p}_{m,i} = p_{m,i} + 2\pi \check{f}_{m,i} \Delta_T(m)$$

$$\Delta_{\varphi_o}^*(m+1) = \{\Delta_{\varphi_o}^*(m) + 2\pi F_0 D(Z_T(m) Z_F(m) - 1)\} \tilde{\text{mod}} 2\pi$$

where $\Delta_{\varphi_o}^*(0) = 0$, and the scale factors $Z_T(m)$ and $Z_F(m)$ are not necessarily constant. Further we define:

$$a \tilde{\text{mod}} b = \begin{cases} a \bmod b & \text{if } a \bmod b < b/2 \\ a \bmod b - b & \text{else} \end{cases}$$

Scaling the frequency axis does not only modify the pitch, but also moves the formants and scales the speech bandwidth. These side-effects are inevitably connected with the frequency scaling and have to be compensated. The original formants can easily be reconstructed by modifying the amplitude of the scaled spectral components using

$$\check{a}_{m,i} = \frac{S_m(\check{f}_{m,i})}{S_m(f_{m,i})} a_{m,i} = \frac{S_m(Z_F f_{m,i})}{S_m(f_{m,i})} a_{m,i}. \quad (3)$$

The spectral envelope S_m of frame m can be determined by linear prediction or by cepstrally smoothing the spectrum. Perceptually much better results, however, have been attained by cosine interpolation of the true log amplitude spectrum.

For $Z_F > 1$ the frequency scaling can cause aliasing which has to be suppressed. On the other hand, for $Z_F < 1$ the bandwidth is reduced. This is particularly irritating for time-varying Z_F , e.g. in a TTS system. Usually this is omitted by keeping the original high frequency components in the bandwidth gap. A perceptually somewhat better approach is to compensate the reduced bandwidth by scaling the components with $\check{f}_i(m) > Z_F(m) F_s/2$ using

$$\check{f}_i(m) = F_s/2 - Z_F(m) (F_s/2 - f_i(m)). \quad (4)$$

An example of a time- and pitch-scaled signal is shown in Figure 5. Additionally, an acoustical demonstration is given in [SOUND A77S01.WAV] using all combinations of time and pitch scaling factors of 0.8, 1.0, and 1.25.

5. CONCLUSIONS

Our investigations have shown that by means of the Fourier analysis/synthesis technique it is possible to modify the prosodic parameters of speech signals with nearly no loss of

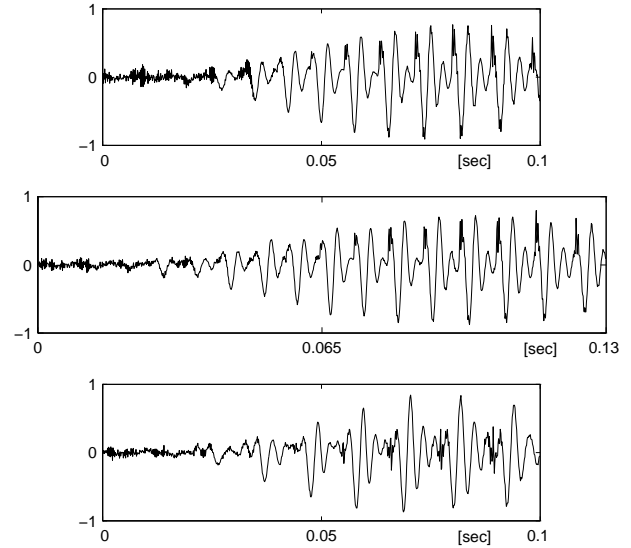


Figure 5: The original signal (top), scaled to 130 % time (middle), and to 73 % frequency (bottom).

quality. The requirements are, however, an accurate estimate of the true short-time spectrum, of the spectral envelope, and of the F_0 . For the scaling operation in the frequency domain it is very important to omit long time shifts.

The presented method is well-suited for TTS systems based on diphone concatenation, where the scaling factors are not constant.

6. REFERENCES

1. B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2):637–655, August 1971.
2. F. Charpentier and E. Moulines. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In *Proc. of the Eurospeech*, pages 13–19. European Speech Communication Association, 1989.
3. R. E. Crochiere. A weighted overlap-add method of short-time Fourier analysis/synthesis. *IEEE Transactions on ASSP*, 28(1):99–102, 1980.
4. C. Hamon. Procéd  et dispositif de synth se de la parole par addition-recouvrement de formes d’onde. Brevet europ en EP 0 363 233 B1 (titulaire: France T l com) dans Bulletin 94/48, Office europ en des brevets, nov. 1994.
5. R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on ASSP*, 34(4):744–754, August 1986.
6. B. Pfister. *Prosodische Modifikation von Sprachsegmenten f r die konkatentative Sprachsynthese*. Diss. Nr. 11331, TIK-Schriftenreihe Nr. 11 (ISBN 3 7281 2316 1), ETH Z rich, M rz 1996.