SYNCHRONIZATION OF SPEECH FRAMES BASED ON PHASE DATA WITH APPLICATION TO CONCATENATIVE SPEECH SYNTHESIS

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

Yannis Stylianou

AT&T Labs-Research, Shannon Laboratories, 180 Park Ave, Florham Park, NJ 07932-0971
yannis@research.att.com
http://www.research.att.com/projects/tts

10.21437/Eurospeech.1999-512

ABSTRACT

Synchronization of speech frames is an important issue in a concatenative speech synthesis system. In terms of signal processing this is translated in removing linear phase mismatches between concatenated speech frames. This paper presents two novel approaches to the problem of synchronization of speech frames with an application to concatenative speech synthesis. Both methods are based on a processing of phase spectra without decreasing the quality of the input speech, in contrast to previously proposed methods. The first method is based on the notion of center of gravity and the second on differentiated phase data. The proposed methods have been tested with the Harmonic plus Noise Model, HNM, in the context of Text-to-Speech synthesis. The resulting synthetic speech is free of linear phase mismatches.

1. INTRODUCTION

One important issue in concatenative speech synthesis is that of synchronization of speech frames, or, in other words, inter-frame coherence. Inter-frame incoherence in the synthetic speech signal cause misalignments of the glottal closure instants which is perceived as a "garbled" speech quality by listeners. In terms of signal processing, inter-frame incoherence means mismatches of the linear phase of the concatenated speech frames. Various strategies have been proposed for elimination of phase mismatches during acoustic unit concatenation. These include: 1) Marking of glottal closure instants (pitch marking) in the speech database [1]. This technique is a time-consuming task and not a completely automated process. Therefore, this method is less suitable for marking big speech databases. 2) Resynthesis of the voiced segments of a speech database by constraining the pitch and phase to be constant [2]. This artificial processing decreases the quality of speech. 3) Replacing the original phase with zero or minimum phase [3]. This approach produces low quality speech. 4) Estimation of the so-called pitch onset time [4], a process which is not always successful [5]. 5) Estimation of phase offset using cross correlation functions [6]; this approach increases the complexity of the synthesizer.

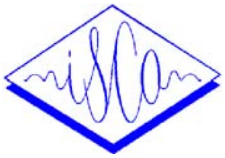
This paper presents two novel ways of removing inter-frames incoherence. Voiced frames are extracted from speech signals; each frame has a duration of two local pitch periods *regardless* where the glottal closure instant is. These frames can be synchronized independently if we decide apriori on a common synchronization point. The first proposed method is based on the notion of center of gravity applied to speech signals [7], where the common synchronization point is decided to be the center of gravity of the speech frames. The second proposed method is based on the differentiated phase data extracted from a speech frame; the common synchronization point for the second method is decided to be the center of the analysis window. We show that this frequency domain approach can be easily transformed into a time domain approach without the transformation of the speech signal from the time domain to the frequency domain. Therefore, the method does not require phase unwrapping.

There are two important advantages of using the proposed techniques. First, estimating the linear phase component from speech frames that are going to be concatenated and then subtracting (explicitly as in the first methods, or implicitly as in the second method) this component from their phase spectra, the quality of the speech signals is not degraded. Second, because both methods achieve the synchronization of the speech frames using only the phase spectrum of the frame that is processed, the synchronization can be carried out during the analysis of the database which is an off-line process. This is mainly due to the fact that there is an apriori agreed-upon synchronization point.

The paper is organized as follows. A review of the notion of center of gravity for speech signals is given first followed by a description of the proposed method. Section 3 describes the second method and compares it with the first method. The application of both methods to the problem of removing linear phase mismatches is presented in Section 4. In order to support our conclusions, Section 5 presents results from the application of both methods within the context of the Text-to-Speech synthesis system of AT&T [8].

2. METHOD I: CENTER OF GRAVITY

This section briefly presents the notion of center of gravity, its connection to the delay of signals and its application to speech. An extended analysis of the subject can be found in [7].



The center of gravity, η, of $f(t)$ is given by:

$$\eta = \frac{m_1}{m_0} \tag{1}$$

where m_n is the nth moment of $f(t)$:

$$m_n = \int_{-\infty}^{\infty} t^n \, f(t) dt \tag{2}$$

With $F(\omega) = A(\omega)\, e^{j\,\phi(\omega)}$ the Fourier transform of signal $f(t)$ and $F^{(n)}(0)$ denoting the nth derivative of Fourier transform of $f(t)$ at the origin, $\omega = 0$, we can show that [9]:

$$F^{(n)}(0) = (-j)^n \, m_n \tag{3}$$

From Eqs. (1) and (3), the center of gravity of $f(t)$ is given by:

$$\eta = \frac{j\, F^{(1)}(0)}{F(0)} \tag{4}$$

where

$$F(0) = \int_{-\infty}^{\infty} f(t)\, dt \tag{5}$$

is the area, m_0, of $f(t)$ and $F^{(1)}(0)$, assuming that $f(t)$ is real, is given by:

$$F^{(1)}(0) = j\, A(0)\, \phi^{(1)}(0) \tag{6}$$

From Eqs. (4) and (6) it follows that:

$$\eta = -\phi^{(1)}(0) \tag{7}$$

This means that the center of gravity, η, of $f(t)$ is a function only of the first derivative of the phase spectrum at the origin ($\omega = 0$).

Let us consider as an example the delta function, $\delta(t)$, and a version of it delayed by t_0, $\delta(t - t_0)$. From Eq.(7) it follows that the center of gravity for the first signal is zero while the center of gravity of the second signal is:

$$\eta = -\phi^{(1)}(0) = t_0 \tag{8}$$

This means that if a signal is delayed by an amount t_0, its center of gravity will be delayed by the same amount. Thus, if the signal has its center of gravity at the origin (as the delta function, $\delta(t)$, does) and its Fourier transform is computed at a distance of t_0 away from the origin then the derivative of the phase at the origin ($\omega = 0$), $\phi^{(1)}(0)$, will be equal to the delay t_0.

Signals with significant values around a time instant t_0, $|t - t_0| \leq d$, while outside of this interval having only insignificant (e.g., zeros) values relative to the values inside the interval, have similar properties as the simple delta function. It turns out that the Linear Prediction, LP, residual signals of voiced speech belong to this class of signals, as long as only one pitch period is considered. Using properties of linear systems we have shown in [7] that the centers of gravity of the speech signal and of the LP residual signal approximately coincide. It follows then that the first derivative, $\phi_s^{(1)}(\omega)$, of the phase function of the speech signal at the origin ($\omega = 0$) is approximately equal to the first derivative, $\phi_r^{(1)}(\omega)$, of the phase function of the LP residual signal at the same point:

$$\phi_r^{(1)}(0) \simeq \phi_s^{(1)}(0) \tag{9}$$

From the above discussion it also follows that the glottal closure instant (GCI) for these signals is very close to their center of gravity; therefore, the estimation of the center of gravity of the speech signal may be used for the detection of GCIs. This is true for most voiced sounds, with the exception of nasals (because in this case the vocal tract is not well represented by a minimum phase filter). Hence, marking the glottal closure instants (like in TD-PSOLA), is not a safe way of removing linear phase mismatches (see discussion in [10] page 258). Replacing GCIs by the center of gravity of speech signals, solves the above problem and, in addition, makes the center of gravity an alternative reference point for voiced frames that can be determined much more robustly than the GCIs.

Based on Eq. (9), and on the fact that the speech signal is a real signal ($\phi(0) = 0$), plus the assumption that the excitation signal for voiced speech can be approximated with a train of impulses:

$$\delta(t - k\, T_0), \quad k = -\infty, \cdots, +\infty \tag{10}$$

we have further shown that the derivative of the phase of the speech signal at the origin is given by [7]:

$$\phi^{(1)}(0) = \frac{\phi(\omega_0)}{\omega_0} \tag{11}$$

where $\omega_0 = 2\pi / T_0$.

If $\phi(\omega)$ denotes the phase spectrum of a speech frame of two pitch periods, measured at time $t = t_0$, and $\theta(\omega)$ denote the unknown phase at the center of gravity, η, of the speech frame ($\theta^{(1)}(0) = 0$), then,

$$t_0 = -\phi^{(1)}(0) \tag{12}$$

since

$$\theta(\omega) = \phi(\omega) + \omega\, t_0 \tag{13}$$

Then, from Eqs. (13), (12) and (11) it follows that the estimated phase, $\hat{\theta}(\omega)$, at the frequency samples $k\,\omega_0$ is given by:

$$\hat{\theta}(k\,\omega_0) = \phi(k\,\omega_0) - k\,\phi(\omega_0) \tag{14}$$

Thereafter, we will refer to Eq. (14) as the correction of the measured phase $\phi(\omega)$ based on the notion of the Center of Gravity.

3. METHOD II: DIFFERENTIATING THE SPECTRUM

Going back to the example of the delayed (by t_0) delta function, $\delta(t - t_0)$, it is easy to see that the delay t_0 can be retrieved from the spectrum information using a *differentiating* function of the spectrum. Indeed, since $\delta(t - t_0) \stackrel{\mathcal{F}}{\longleftrightarrow} e^{-j\,\omega\,t_0}$ then:

$$arg\{e^{-j\,\omega\,t_0}\, e^{j\,(\omega+\Delta\omega)\,t_0}\} = \Delta\omega\, t_0 \tag{15}$$

In case $\Delta\omega = 2\pi/N$ is the sampling frequency of the spectrum:

$$t_0 = \frac{N}{2\pi}\, arg\{e^{-j\,\omega\,t_0}\, e^{j\,(\omega+\Delta\omega)\,t_0}\} \tag{16}$$

The above result can be easily generalized. If $F(\omega)$ and $f(t)$ form a Fourier-integral pair, then an estimator of the

delay t_0 of $f(t)$ from the center of the analysis window is the argument of the integral:

$$\int_{-\infty}^{\infty} F(\omega)\, F^*(\omega')d\omega \qquad (17)$$

where $F^*(\omega)$ denotes the conjugate of $F(\omega)$, and $\omega' = \omega + \Delta\omega$. It is easy to show that the above integral is equivalent to a time domain integral, thus, avoiding the computation of Fourier transform of the signal and any necessary phase unwrapping:

$$\int_{-\infty}^{\infty} F(\omega)\, F^*(\omega')d\omega = \int_{-\infty}^{\infty} f^2(t)\, e^{j\,\Delta\omega t}\, dt \qquad (18)$$

4. APPLICATION TO SPEECH SYNTHESIS

Let $s_w[n]$ denote a voiced speech frame weighted by a window $w[n]$ with a length of two pitch periods ($2\,T_0$), and $\phi(k\,\omega_0)$ the estimated phase at multiples of fundamental frequency $\omega_0 = 2\pi/T_0$. The phase may be estimated by a simple peak picking of the spectrum of $s_w[n]$. In this paper, the phase is estimated by minimizing a weighted time-domain least-squares criterion [11]:

$$\epsilon = \sum_{n=-T_0}^{T_0} [s_w[n] - w[n]\, s_h[n]]^2 \qquad (19)$$

where $s_h[n]$ is a harmonic signal to estimate and T_0 is the local fundamental period.

Correcting the estimated phase spectrum using the method of Center of Gravity is a straightforward process by applying Eq. (14). This results in moving the center of the analysis window, $w[n]$, to the center of gravity of $s_w[n]$, independently of the initial position of the window. For the second proposed method, the delay t_0 is given by:

$$t_0 = \frac{T_0}{2\,\pi} arg \sum_{n=-T_0}^{T_0} s_w^2[n] e^{j\,\omega_0\,n} \qquad (20)$$

and then applying Eq. (13). This moves the "area", $\mathcal{A}$, which encloses the maximum value of the squared windowed signal, $s_w[n]$, to the center of the analysis window. This area becomes very narrow in case that $s_w[n]$ is a perfectly harmonic signal (ideal case). While this area is wider for speech signals, the method achieves the synchronization of speech frames and it is not sensitive to errors ($< 10\%$) in the estimation of fundamental frequency.
In most of the cases the two methods agree in the estimation of the delay, since the center of gravity of $s_w[n]$ coincides with the area $\mathcal{A}$. However, there are some differencies for sounds like nasals.

Figure 1 shows an example of phase correction using a speech signal. The left column of the figure shows the different position of the analysis window before phase correction while the right column shows it after phase correction with the method of the center of gravity (solid line) and with the differentiating spectrum method (dashed line). The frames after phase correction are aligned. As the figure indicates the analysis window is two pitch periods long.

Proceeding in a similar manner for all voiced frames results in automatically aligning all frames at their center of gravity. Thus, synchronization of frames is assured when speech frames are concatenated for text-to-speech synthesis. The important point to note is that the synchronization of frames is achieved without estimation of glottal closure instants and independently of the frames that are concatenated.

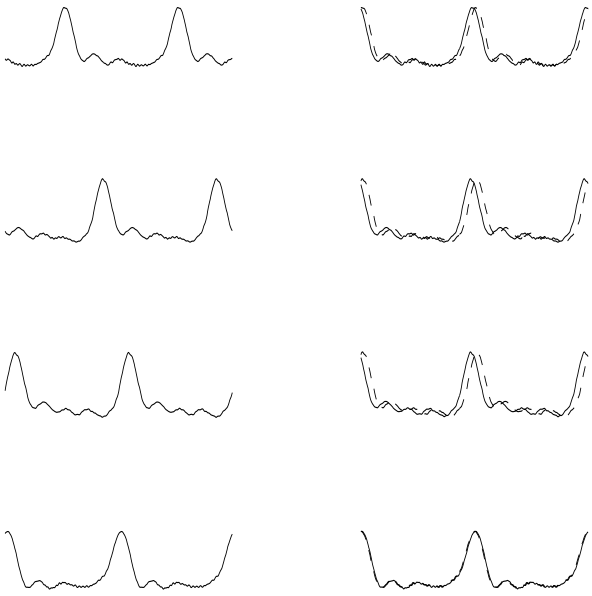


Figure 1: Phase correction. Position of analysis window before phase correction (left) and after phase correction (right) with the method of center of gravity (solid line) and the method of differentiating spectrum (dashed line). Signals are shown without the weighting function $w[n]$.

5. RESULTS AND DISCUSSION

From the above presentation of the two proposed methods, it is clear that they can be successfully applied in removing phase mismatch (incoherence of speech frames) in concatenative speech synthesis. As both methods can be used off-line (during the analysis of the speech database) they can be useful for many different speech synthesizers (e.g., TD-PSOLA, MBROLA and HNM). In HNM, phase correction is a straightforward process. The analysis windows are placed in a pitch synchronous way regardless of where glottal closure instants are. Phases are estimated by minimizing a criterion similar to the one in Eq. (19) [11] and are corrected using Eq. (14) or Eq. (13) after estimated the delay t_0 using Eq. (20). Fig. 2 shows four signal segments in the vicinity of their concatenation points. The segments have been extracted from a speech synthesis example using HNM for concatenation of diphones. The left column of the figure shows concatenation of the segments without any prior phase correction, while the right column shows the same segments with an off-line phase correction based on the center of gravity method (similar results are obtained using the method based on the differentiation of spectrum). It is clear that the proposed methods efficiently remove any phase mismatch from the speech segments allowing a good synchronization of the

frames across the concatenation point. Both methods

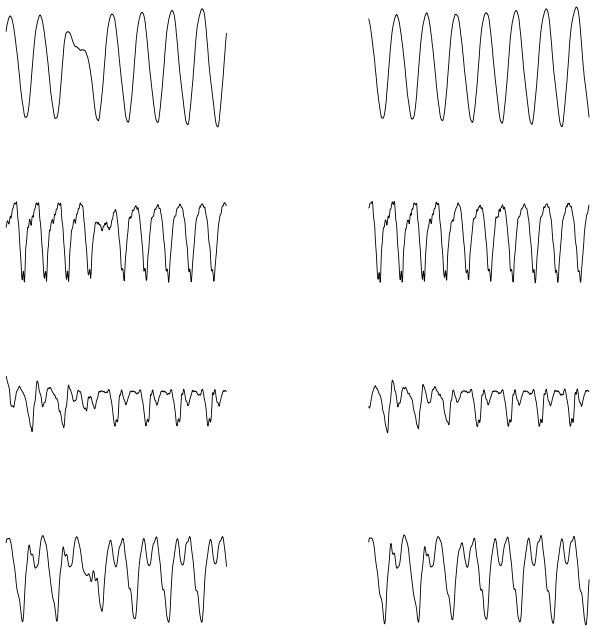


Figure 2: Example from speech synthesis (using HNM) of the text: *I'm waiting for my pear tree to bear fruit.* Left: Concatenation without applying any phase correction algorithm (e.g., cross-correlation). Right: Concatenation after phase correction using the center of gravity method.

have been applied in the context of Harmonic plus Noise Model, HNM, for speech synthesis.

AT&T's Next-Generation Text-to-Speech synthesis system [8] is based on unit concatenation (half-phonemes, diphones, or longer units). Given an input text and a desired prosody for this text a unit selection algorithm selects an optimum set of units. In this context, a large speech database has to be used in order to provide the unit selection algorithm with many instances of a unit. Currently, about two hours of recording of a female speaker are used as database. Applying the proposed algorithms, the acoustic units are concatenated without any phase problem. The methods have also been used for speech synthesis based on concatenation of diphones with other voices as well. The test corpus included eight professional American male speakers, one male voice for British English, a male voice for French and five other female voices for American English. For all these voices and databases the proposed methods completely remove any phase mismatch between voiced frames.

The synchronization of speech frames is also important for the representation (or coding) of speech signals. For instance, HNM performs a time-varying harmonic plus modulated noise decomposition of the speech signal. Knowing the position of the harmonic part, the noise part can be synchronized accordingly. This makes the noise to be perceptually integrated with the harmonic part, improving the quality of the speech signal synthesized by the model. The proposed synchronization techniques can be also used in reducing the complexity of speech coders where speech frames have to be synchronized e.g., the Waveform Interpolation (WI) [12] coding system.

6. CONCLUSION

In this paper we presented two methods for the synchronization of speech frames with an application to concatenative speech synthesis. The proposed methods have been tested on large speech databases for male and female speakers. No errors have been observed in removing phase mismatch during synthesis. Contrary to previously reported methods that have been proposed as solutions to the phase mismatch problem, the proposed methods are simple and efficient. Moreover, they don't require additional complexity during synthesis and they don't modify the quality of the speech signal.

7. REFERENCES

[1] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, Dec 1990.

[2] T. Dutoit and H. Leich, "Text-To-Speech synthesis based on a MBE re-synthesis of the segments database," *Speech Communication*, vol. 13, pp. 435–440, 1993.

[3] M. Crespo, P. Velasco, L. Serrano, and J. Sardina, "On the Use of a Sinusoidal Model for Speech Synthesis in Text-to-Speech," in *Progress in Speech Synthesis*, pp. 57–70, Springer, 1996.

[4] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, Aug 1986.

[5] M. W. Macon, *Speech Synthesis Based on Sinusoidal Modeling*. PhD thesis, Georgia Institute of Technology, Oct 1996.

[6] Y. Stylianou, T. Dutoit, and J. Schroeter, "Diphone Concatenation using a Harmonic plus Noise Model of Speech," *Proc. EUROSPEECH*, pp. 613–616, 1997.

[7] Y. Stylianou, "Removing phase mismatches in concatenative speech synthesis.," *Third ESCA Speech Synthesis Workshop*, pp. 267–272, Nov. 1998.

[8] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS System.," *137th meeting of the Acoustical Society of America*, 1999. http://www.research.att.com/projects/tts.

[9] A. Papoulis, *Signal analysis*. New York: McGraw-Hill, 1984.

[10] T. Dutoit, *An introduction to text-to-speech synthesis*. The Netherlands: Kluwer Academic Publishers, 1997.

[11] Y. Stylianou, J. Laroche, and E. Moulines, "High-Quality Speech Modification based on a Harmonic + Noise Model.," *Proc. EUROSPEECH*, pp. 451–454, 1995.

[12] W. Bastiaan Kleijn and J. Haagen, "Waveform Interpolation for Coding and Synthesis," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), ch. 5, pp. 175–207, Marcel Dekker, 1991.