AUTOMATIC POST-SYNCHRONIZATION OF SPEECH UTTERANCES

Werner VERHELST

Vrije Universiteit Brussel, Faculty of Applied Science
Dept. of Electronics and Signal Processing (ETRO)
Pleinlaan 2, B-1050 Brussels, Belgium
E-mail: wverhels@vnet3.vub.ac.be

ABSTRACT

The paper considers a prototype for automatic post-synchronization that consists of two basic components. As a first step, dynamic time warping is applied to compute the time-correspondence between an original utterance and an utterance that serves as the timing reference signal. In a second step, a time-scaling algorithm modifies the time structure of the original utterance accordingly. Informal diagnostic evaluation has shown that good results are obtained if the similarity between the acoustic-phonetic contents of the utterances is high. Possible ways for improving robustness against acoustic-phonetic differences, such as those that result from different coarticulation, are suggested.

1. INTRODUCTION

In recent years, techniques have been developed that allow prosodic modifications of speech without serious quality degradation (see, e.g., [1]). With these techniques, a number of applications can be developed that were previously unrealistic because they require a processing that is transparent to the listener.

In this paper, we describe a prototype system for automatic post-synchronization of speech utterances. The goal of the system is to modify the timing structure of an original speech utterance in such a way as to synchronize it with a second utterance, which has the same textual content and could have been produced by the same or by a different speaker.

Possible uses for such a system can be found for example in the audio and audio-for-video industries: automatic post-synchronization could be applied in the audio industry to artificially create or improve choral singing or to enrich the timbre of a single voice by mixing together several synchronized recordings; in audio-for-video production, it could be applied for dubbing material with another voice than the original one (which should be useful for commercials) or for dubbing outdoor recordings with studio material (synchronization with the original speech soundtrack could be used to improve synchronicity with the image).

Section 2 of this paper describes the design of the prototype. The experiments that were performed are described in Section 3. A discussion of the results and suggestions for further developments are given in Section 4 before concluding the paper in Section 5. A number of sound examples are also included on the Eurospeech'97 CD-ROM to illustrate the performance (see section 3.3).

2. SYSTEM PROTOTYPE

The proposed system uses a two-step procedure:

1. Dynamic Time Warping (DTW) is used to compute the timing relationship (time-warping path) between the original utterance and the utterance that serves as timing reference.
2. The original signal is time-scaled in accordance with the time-warping path, such that the result is synchronous with the timing reference.

Fig. 1 illustrates the timing relations between an original and a timing reference, and shows good synchronization between the time-scaled result and the timing reference (relations between panels are shown explicitly for three different instances).

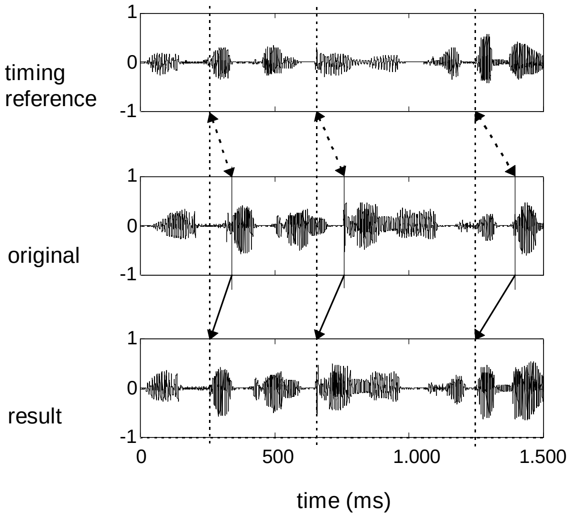

Figure 1. *After analyzing its relation to the timing reference, the original is time-warped to cancel the existing differences.*

2.1. Dynamic Time Warping

A short-time LPC analysis is performed on both the timing reference and on the original signal. A matrix is constructed with elements *d(j,i)* , *j=1.. J*, *i=1.. I* which are equal to the spectral distances between frames *j* of the timing reference and frames *i* of the original, and where *J* and *I* represent the number of frames in the respective signals. The time-warping path is obtained as the path (j_k , i_k) that minimises the accumulated distance *D*

$$D = \sum_{k=1}^{N} d\left(j_k , i_k\right)$$

subject to the constraints

$$\begin{aligned} & j_1 = i_1 = 1 \\ & j_N = J; i_N = I \\ & \left(j_{k-1}, i_{k-1}\right) \in \left\{\left(j_k - 1, i_k\right), \left(j_k - 1, i_k - 1\right), \left(j_k , i_k - 1\right)\right\} \end{aligned}$$

A bandpass liftered cepstral distance measure [2] is used to compute the spectral differences between the individual frames:

$$\begin{aligned} & d\left(j_k , i_k\right) = \sum_{n=1}^{M} w(n)\left(c_{j_k}(n) - c_{i_k}(n)\right)^2 \\ & w(n) = 1 + 6\sin(\pi n / M) \\ & M = \left\lceil 12 f_s / 6600 \right\rceil \end{aligned}$$

where f_s is the sampling frequency in Hz, and c_{j_k} and c_{i_k} represent the LPC cepstral vectors for frames j_k and i_k of the timing reference and of the original signal, respectively. (Like in [3], the lifter length is proportional to the sampling frequency and the gain term (*n=0*) is omitted.)

Dynamic Time Warping has been extensively studied for speech recognition and more sophisticated forms for the functional *D* and for the constraints have been proposed which improve recognition scores in a number of systems [4]. However, in a previous study [3] we could not clearly find them advantageous as far as the accuracy of the time-warping path was concerned (they often introduced inaccuracies when differences between acoustic realisations occurred, e.g., when one of the utterances contained a breathing pause). Therefore we opted to use the basic version of DTW for this prototype, as described above.

2.2. Time-Scaling Algorithm

The time-scaling algorithm is used to produce a time-scaled version of the original in accordance with the time-warping path. Therefore, it should be capable of high-quality time-scaling with arbitrarily time-varying factors. We used WSOLA for this purpose ([1],[5]).

As illustrated in Fig. 2, WSOLA constructs the output signal *y(n)* by overlap-adding windowed segments from the original *x(n)*. We used a hanning window with length 2*L*=15*ms* and 50% overlap between successive synthesis segments (i.e., L_k = *kL*).

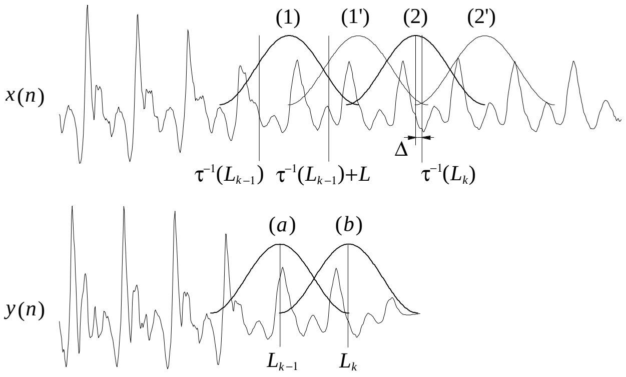

Figure 2. *Illustration of WSOLA time-scaling.*

First we construct the desired time-warping function *τ(n)* by linear interpolation between successive points (j_k, i_k) of the time-warping path. Then, referring to the illustration in Fig. 2 and proceeding in a left-to-right fashion, assume we previously excised segment (1) from the input signal and overlap-added it to the output in position (a) (at time instant L_{k-1} = (*k*–1).*L*). We thus need to find a next segment (2), located somewhere about time instant τ^{-1}(*k.L*) in the original signal, that will produce a natural continuation of the output signal when added in position (b), i.e., without producing pitch discontinuities or phase jumps. As (1') would overlap-add with (1) in a natural way to form a portion of the original input signal, WSOLA chooses (b) such that it resembles (1') as closely as possible and is located within a prescribed tolerance interval [-$\Delta_{\max}$.. $\Delta_{\max}$] around τ^{-1}(*k.L*) in the input wave. The precise position of this best segment (2) is found by maximising a similarity measure (the cross-AMDF in this case) between the sample sequence underlying (1') and the input signal. After excising (2) and adding it in position (b), WSOLA proceeds to the next output segment, where (2') now plays the same role as (1') in the previous step.

3. EVALUATION

3.1. Experimental Procedure

The performance of the system depends on the quality of the processed speech and on the accuracy of the time alignment. Since neither of these can be easily quantified, a qualitative evaluation was used (DTW does not qualify as an independent measure for time-alignment in this case as it is used by the system). Fortunately, in this case qualitative evaluations can be performed relatively easily. As WSOLA uses a form of automatic signal editing that preserves much of the quality and acoustic detail of the original, the processed sound quality can be estimated from the amount and prominence of audible distortions in the result in comparison to those in the original signal and in the timing reference. The accuracy of the time alignment is made clear when the synchronization result is mixed with the timing reference signal and played simultaneously: all relevant time alignment errors are easily heard-out.

The prototype was implemented as an off-line system on a PC and a number of sentence long utterances (around three seconds in length) were recorded by a number of speakers (male, female and child) using a plug-in sound-card. When needed for testing the effect of noisy timing reference signals, noise signals were added to the recordings. So far the evaluation has been performed by the author in informal diagnostic listening sessions using equipment that varied from small multimedia loudspeakers to good quality consumer headphones.

3.2. Experimental Results

The time scaling accuracy of WSOLA appeared to be sufficiently high for this application. Because it works with segments from the original utterance, WSOLA requires a tolerance [$-\Delta_{max}$.. Δ_{max}] on the time-warping path to ensure sufficient continuity between successive segments. With 15 *ms* segments and accurate timing to within a tolerance of [-7*ms*.. +7*ms*], signal continuity was ensured without introducing noticeable time-misalignments. Thus, the timing precision of WSOLA was judged high enough for this kind of applications.

Perceived distortions in the output could often be traced to some event in the time-warping path, but could not always be considered to be due to an error in the path. In general, one could say that results with the current system are good when the timing reference and the original are acoustic-phonetically sufficiently close since distortions could often be attributed to different types of acoustic-phonetic differences:

- Impossible insertions. When a phone is pronounced carefully in the timing reference, but is absent or heavily coarticulated and short-lived in the original, a correct synchronization would require that a few transient frames of the original be heavily time-stretched. As a time-scaling algorithm does not adapt the spectral characteristics of the speech segments accordingly, the result could be perceived to be distorted.
- Incomplete deletions. In the reverse situation too, problems can occur without any of the system components being responsible. When strong coarticulation occurs in the timing reference but not in the original, the time warping path should specify substantial shortening for the implicated phones from the original. The transitions into and away from these phones in the time-scaled result could then seem too slow and too carefully produced, resulting in a synchronized utterance that appears to be missing some part.
- Incompatible substitutions. Some allophones could have different acoustic realisations in the original and the timing reference. If the inherent length of these different realisations differs, the system will produce the acoustic variant from the original with the duration of the timing reference, again leading to perceive a distortion, even in the absence of time warping or time scaling errors. A comparable situation might occur when distortions or noises exist in the original: they could become more prominent after time-scaling (an example occurs in the first set of illustrations on the CD-ROM at the end of original and synchronized utterances).

These problems are characteristic for the application in that they occur without that either the time-warping or the time-scaling procedure can be held responsible for the distortion. In section 4, we discuss some approaches that could possibly help in dealing with acoustic-phonetic differences in this type of application.

When noise is added to the timing reference signal or when the speaker of the timing reference is different from that of the original, the distortion problems appear to occur with increased frequency compared to the single speaker clean-speech situation, which is consistent with the hypothesis that problems are related to acoustic-phonetic differences.

3.3. Audio Demonstration (CD-ROM)

This subsection presents the audio demonstrations that were produced for the Eurospeech'97 CD-ROM. (With the CD-ROM, the utterances should play when activating the corresponding links).

A first set of utterances illustrates automatic synchronization in the single speaker clean speech situation:

- [original A1083S01.WAV]
- [synchronized result A1083S02.WAV]
- [timing reference A1083S03.WAV]

(The first portions of the oscillograms of these utterances are shown in Fig. 1.)

The second set illustrates the effect of a noisy timing reference signal (this situation could occur in practice when dubbing outdoor recordings for example):

- [original A1083S03.WAV]
- [synchronized result A1083S04.WAV]
- [timing reference A1083S05.WAV]

The next set of utterances illustrates a situation where different speakers are used:

- [original A1083S06.WAV]
- [synchronized result A1083S07.WAV]
- [timing reference A1083S08.WAV]

To illustrate the accuracy of the synchronization:

- [play A1083S09.WAV] a mix of the timing reference and the synchronized result which demonstrates good synchronization
- [and compare A1083S10.WAV] to a mix of the timing reference and the uniformly time scaled original (scaled to give it the same duration as the timing reference).

4. DISCUSSION

Automatic post-synchronization with natural sounding results would seem possible with the present basic implementation, provided that the acoustic-phonetic similarity between the original and the timing reference is sufficiently high. In practice there can be more variability between different realisations, especially if different speakers are involved or with spontaneous speaking styles. Therefore, the robustness of the system against local acoustic and phonetic differences should be improved.

Since it is impossible to insert a missing phone by time-stretching, this type of problem could be resolved by imposing local slope constraints [4] on the warping path. Unlike the common practice in speech recognition applications, however, it could be advisable not to use symmetrical slope constraints in this case, as one should still be able to eliminate insertions like breathing pauses from the original in post-synchronization applications. Therefore, it would seem appropriate to develop asymmetric slope constraints for these applications, restricting only the allowed time expansion factor. Hopefully, this would lead to stretching the surrounding phones from the original such as to cover the duration of the extra phones in the timing reference and would sound more natural.

The case of incomplete deletions, and perhaps to some degree of incompatible substitutions, might require constraining the rate of change of the slope of the warping path (viz., the rate of change of the relative speaking rate of one utterance compared to the other). This could again lead to more gradual time-scaling and more smoothly sounding results. Furthermore, it seems reasonable that in natural speech the speaking rate would be relatively slowly time-varying as well. Perhaps a non-linear smoothing operation, applied on the warping path in a postprocessing step, could be a suitable way of implementing this type of constraint.

5. CONCLUSION

The idea of automatic post-synchronization has been explored using a basic implementation of the system. The observations made indicate that the development of such an application could be viable in practice. To achieve this, further work should mainly concentrate on the development of a proper user-interface and on improved strategies for computing more natural sounding time-warping paths.

REFERENCES

[1] E. Moulines, W. Verhelst, “Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech”, CH 15 in: W.B. Kleijn, K. K. Paliwal (eds.), *Speech Coding and Synthesis*, Elsevier 1995.

[2] L. R. Rabiner, F. K. Soong, “Single-Frame Vowel Recognition Using Vector Quantization with Several Distance Measures”, AT&T Technical Journal, vol. 64, nr. 10, pp. 2319-2330, 1985.

[3] W. Verhelst, M. Borger, “Intra-Speaker Transplantation of Speech Characteristics”, proceedings Eurospeech’91, pp. 1319-1322, 1991.

[4] J. R. Deller Jr., J. G. Proakis, J. H. L. Hansen, *Discrete-Time Processing of Speech Signals* (CH 11), Macmillan 1993.

[5] W. Verhelst, M. Roelands, “An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech”, proceedings ICASSP-93, vol. II, pp. 554-557, 1993.