Speech Synthesis for a Specific Speaker Based on a Labeled Speech Database

R. Hoory†
Dept. of Electrical Engineering
Technion, Israel Institute of Technology
Haifa 32000, Israel

D. Chazan
HRL. Audio/Video Technologies group
IBM Israel – Science and Technology
Haifa 31905, Israel

Abstract
This paper proposes a new text-to-speech synthesis technique, for producing continuous, natural sounding speech of a specific speaker. The synthesis technique is based on selecting short speech frames from a phoneme-labeled speech database. The selection procedure involves minimization of a distortion criterion, by a dynamic programming algorithm. The proposed scheme is more flexible than many existing schemes using fixed speech segments, such as diphones. It results in a more natural synthesized speech. An efficient speech representation is used to express simply and accurately the spectral continuity of speech. A further improvement in the database search mechanism and in database size was obtained by sectioning the speech phonemes into “steady-states” and “transitions”. The resulting synthesized speech quality, is satisfactory and indeed preserves the natural voice of the speaker.

1 Introduction
In concatenative speech synthesis methods, acoustic segments, previously extracted from natural speech are assembled, to form new words and sentences. Various type of segments, such as demisylables [1], diphones [8], and allophones (or pseudo-phonemes) [7] are commonly used. With the use of diphones to characterize transition sounds and coarticulation effects, the main difficulty is the maintenance of spectral and phase continuity. One way to treat this problem is to improve the pre-synthesis extraction of the synthesis units. For the case of diphone synthesis, Kaelin [4] proposed a methodology of locating the diphone boundaries, where the spectral stability is maximal and the coarticulation minimal. Duttoit [2] produces more consistent diphones by energy and pitch normalization of the database. This is applied to a time domain PSOLA type scheme, which concatenates diphone segments modified to have the correct pitch and duration by time domain manipulation. These approaches while alleviating the problem, do not guarantee spectral and phase continuity of the resulting time signal.

A different approach, which allows greater flexibility in the choice of units, will be considered in this paper. It makes use of flexible synthesis units. Variable length segments are selected during the synthesis process rather than using pre-stored fixed length units. An appropriate distortion criterion allows for an optimal selection of the database speech segments, so that the spectral matching between them would be maximal. The database may consist of arbitrary sequence of sentences, uttered by the speaker which should be phonetically varied. A good selection criterion guarantees a constructive use of the entire database, and a quality which improves as the size of the database grows.

Iwahashi et al. [3] used a flexible unit approach, coupled with a distortion criterion. The drawback of their system is that it limited the synthesis units to phoneme sequences.

The synthesis scheme that is described below makes use of synthesis units which consist of any number of 8 mSec speech frames, and thus allows greater flexibility. With flexible synthesis units, also matching the phoneme durations of the synthesized speech to the desired ones becomes an easy task.

2 The Speech Representation
The speech representation employed here, is somewhat comparable to the formant representation, yet much more detailed, so that a reasonable quality level is maintained.

The speech analysis is carried out in 8 mSec frames (64 samples, with sampling frequency of 8 kHz). During the first stage, the spectral envelope and the pitch period are accurately estimated (see [5, 6]). At the second stage the spectra are divided into a number of non-uniform sub-bands (usually 12 or 16). The width of the sub-bands is determined by a Bark-Scale (see [9]). For each band, the energy and band center are calculated. We can then code each pair of energies, belonging to neighbouring bands, to their sum, and energy-weighted average. This is repeated till we are left with only one energy and a hierarchical tree structure. Figure 1 demonstrates this for 12 sub-bands, leading to a tree of five levels and a total of N = 23 centers.

† Currently with HRL Audio/Video Technologies group, IBM Israel – Science and Technology, Haifa 31905, Israel.
To summarize, each speech frame is represented by the energy $e_i$, a vector $X_i$ of $N$ spectral centers, all normalized to the range of $[-1, 1]$, and the pitch. We shall term $X_i$ spectral vectors. This representation is robust and efficient for compression. It also has the property that its components may be interpreted as frequencies. Hence, a Euclidean metric may adequately serve as a distance measure between spectra.

Using this representation it is possible to reproduce good quality speech.

3 The Speech Database

The database consists of speech sequences of the specific speaker, which were coded according to the representation mentioned above. The speech should be phonetically varied and include all relevant phoneme transitions. It is essential for the speech database to be phoneme-labeled. We use an automatic segmentation scheme, which is consistent with the same speech representation. The segmentation algorithm is beyond the scope of this paper.

4 The Basic Synthesis Scheme

As mentioned, our objective is to select speech frames from the database which will be assembled to produce the synthesized speech. Actually it is sufficient to generate a "synthesis path", defined as a sequence of database indices $\{I_k\}_{k=1}^K$. The synthesis problem is to find an optimal (in the sense described below) synthesis path, under the constraints imposed by the given phonetic transcription and the matching phoneme durations.

The distortion criterion for a synthesis path is based on a measure of spectral matching between database frames. For frames $i$ and $j$ this measure is given by:

$$d(i, j) = \frac{1}{N_v} \sum_{k=-[\frac{N_v}{2}]+1}^{[\frac{N_v}{2}]} \rho_\omega^2(X_{i+k}^{(DB)}, X_{j+k}^{(DB)})$$  \hspace{1cm} (1)$$

where $N_v$ is a positive integer, $X_{i+k}^{(DB)}$ is the spectral vector associated with the $i$th database frame, and $\rho_\omega$ is a weighted Euclidean distance measure. This expression averages the square distance along $N_v$ frames around $i$ and $j$, with an offset of one frame. Using larger $N_v$ is equivalent to matching additional higher order derivatives. This implies a minimal distortion of the time evolution of the spectral components.

The distortion measure of a synthesis path $\{I_k\}_{k=1}^K$ is given by the sum:

$$D(\{I_k\}_{k=1}^K) = \sum_{k=1}^{K-1} d(I_k, I_{k+1})$$  \hspace{1cm} (2)$$

Clearly, for a sequence of successive indices, representing a segment copied "as is" from the database: $D(n, n+1, \ldots, n+K) = 0$.

The optimization of the synthesis path is carried out by minimizing the distortion measure, under some constraints:

$$\min_{\{I_1, I_2, \ldots, I_K\} \in C_k} D(\{I_1, I_2, \ldots, I_K\})$$  \hspace{1cm} (3)$$

where $C_k$ is the set of database indices available for selection at frame $k$. The constraints specified by $C_k$, limit the selection to speech frames with a phonetic context similar to the one required for the synthesized speech. The minimization may be readily carried out, using dynamic programming.

5 Segmenting the Speech to Steady-States and Transitions

Figure 2 presents a two dimensional section through the (frequency diagram) distribution contour lines for all the frames labeled /a/. Next to it is shown a subset of these where spectra lying in transition regions have been rejected. Presumably, these vectors which have been left in, correspond to steady state periods of the phoneme, for which time sequence is not critical. This reduced variance distribution may be represented by a few cluster centroids. These may be obtained by any clustering procedure. The centroid locations are displayed as dots in figure 2 for the two cases.

5.1 Steady-State Extraction Algorithm

In order to determine the locations of the steady-state periods, a spectral stability cost-function is used. The steady-state is the segment, where this function falls beneath a predefined threshold.

The cost function, comparable to the one in [4], weighs two components: The local stability, which is the average distance between the current frame and its neighbors and the global stability, which is the distance to the nearest representative centroid of the phoneme's steady-state.
Initially, no information concerning the locations of the steady-state segments exists. The first iteration extracts the steady-states segments, using the local stability only. Subsequently, each iteration step first updates the clusters for every phoneme, with the steady-state assignments obtained at the previous iteration. This is followed by extraction of the steady-state segments, using the current clusters. The number of centroids should increase from iteration to iteration, as the steady-state representation becomes more and more accurate. It is possible to make use of the LBG vector quantization algorithm with the splitting method.

5.2 The Reorganized Database

Using the internal subdivision of the phonemes, the database speech frames may be categorized to four types: Steady-State (SS), Begin-Transition (TB), End-Transition (TE) and a Transition phoneme (T). Mostly for plosives and short occurrences, Steady-state frames are replaced by the representative centroids, thus achieving about 50% reduction of the database size. The frame type assignment and the location inside the phoneme are also taken into consideration during the synthesis process. For example, at the beginning of the phoneme, TE frames cannot be selected. Combined with the drastic reduction of the amount of steady-state frames, this leads to a much more efficient process, with less room for error. The spectral matching function (1) now depends on the frame types. For steady-state frames, the neighbouring frames are not taken into account.

6 Experimental Results

The synthesis technique described in this paper is designed to produce the spectral components of the speech. Prosodic parameters such as rate, pitch and intensity are supplied as input. During the experimentation, these prosodic parameters were obtained from a reference speech signal with the same phonetic transcription.

For a large enough database, the output of the synthesis procedure usually preserves the voice of the database speaker with reasonable quality and naturalness. Figure 3 demonstrates the synthesis of the Hebrew word /sh-a-d-o-m/ (only the main spectral component is shown). The phoneme locations are shown above the graphs. For the case of synthesis with steady-state quantization (diagram (c)), the steady-state segments determined by the synthesis algorithm, are displayed as well.

7 Conclusions

In this paper, a novel database synthesis technique was introduced. It incorporates both an efficient speech representation and flexible synthesis units. The complexity of the synthesis scheme is relatively low. It currently runs in near real time on an IBM RS/6000 work-station. We believe that a considerable further reduction in running time is feasible.

Further study may be necessary concerning speaker specific prosody. Other applications for which this work has bearing on are: speech compression, speech recognition, speaker recognition, and voice conversion.

Figure 3: The main spectral component for:
(a) The reference signal (b) Signal synthesized with the basic algorithm.
(c) Signal synthesized with steady-state quantization (reorganized database).

References