VOWEL CONVERSION BY PHONETIC SEGMENTATION

Carlos de Obaldía, Udo Zölzer
Helmut Schmidt University
Hamburg, Germany
deobaldia@hsu-hh.de

ABSTRACT

In this paper a system for vowel conversion between different speakers using short-time speech segments is presented. The input speech signal is segmented into period-length speech segments whose fundamental frequency and first two formants are used to find the perceivable vowel-quality. These segments are used to represent a voiced phoneme, i.e. a vowel. The approach relies on pitch-synchronous analysis and uses a modified PSOLA technique for concatenation of the vowel segments. Vowel conversion between speakers is achieved by exchanging the phonetic constituents of a source speaker’s speech waveform in voiced regions of speech whilst preserving prosodic features of the source speaker, thus introducing a method for phonetic segmentation, mapping, and reconstruction of vowels.

1. INTRODUCTION

Applications for segment-based speech analysis and synthesis in engineering and scientific perspectives have been applied for speech morphing algorithms, speech synthesis and coding, and text-to-speech (TTS) systems just to mention a few. In voice conversion (VC) systems, a source speaker’s waveform is converted so that it perceptually resembles the voice of a target speaker whilst retaining linguistic information [1]. Techniques which have a broad success in speech and voice conversion, usually use signal segments as the unit to concatenate during re-synthesis [2].

In linguistics, speech can be modeled as a series of phonemes. Phonology is a branch of linguistics which provides information on the nature of these abstract units. Phonemes follow a particular order to describe words and utterances. Psycholinguistic models of speech production can be abstracted to identify two levels or stages of speech processing: the word (lemma) and the phonological level [3]. In the lemma representation, abstract word properties such as grammatical features are coded. The phonological information, i.e. the form of a word, is coded at the next level of processing [3].

Most of current speech synthesis systems use diphones as a basic selection unit for speech concatenation at the back-end [4, 5]. Diphones are generally constituted by parts of two or more phonetic elements, and contain transitions between phones [6]. However, segments of the speech waveform can also be abstracted to a phonetic level and concatenated attaining to a particular prosody to reproduce a series of phones which form a phoneme [5]. Although using diphones during synthesis may lead to a smooth speech waveform, the use of phone-based analysis and synthesis methods for speech corpora can introduce advantages in context, speed, and consistency [7].

In this paper a phoneme-based vowel conversion system is proposed to convert voiced regions of speech from one speaker (a source speaker) so that it sounds like it was pronounced by another speaker (the target speaker) by using phonetic trajectories instead of supra-segmental models and using signal snippets instead of spectral morphing. Speaker conversion is achieved by synthesising by-rule of the target speaker’s vowel segments using the source speaker’s tonality and intensity.

Perception in humans of a different speaker quality, or timbre, between different speakers is based on differences on the laryngeal and vocal tract resonances [8]. This work focuses only on voiced parts-of-speech, considering that the vocal-tract filter configuration for particular speakers is enough, at least at the phonological level, for a perceptually viable voice conversion.

Since the work is based on the phonetic level of speech production [3], and speech rate is known to contain linguistic information [9], a language independent synthesizer for different speech qualities, for example, could be constructed using the presented method.

2. PROPOSED SYSTEM

A speech segment is any discrete unit that can be identified in a stream of speech and is composed of a series of phonemes. A phoneme is the smallest phonetic contrastive linguistic unit in a language which may bring about a change of meaning. A logical distinction would be to categorize phonemes according to its acoustic nature, for example into voiced and non-voiced sounds.

Voiced phonemes such as vowels are of a periodic nature. Thus a short-time speech segment (STSS) can be extracted from a voiced speech region and characterized as an atomic unit of speech (i.e., a phone) based solely on the information of a particular fundamental period.

Voiced speech is also modeled as the response of the vocal tract to a source excitation, which is generally produced by the opening and closing of the glottal vents in the trachea [10]. The characteristics of such a response are achieved by the amplification of the frequency components of the excitation in the vocal tract.
features of each STSS \( s_j(n) \), the fundamental frequency \( f_0, j \), and the first and second formants \( (F_1, F_2) \) of \( s_j(n) \) are extracted and mapped to the perceived backness and height coordinates of the vowel-space of Fig. 1. The same procedure is used to extract the perceived backness and height measures at each pitch mark of the source speaker’s signal.

Synthesis is performed using PSOLA with the time envelope and the fundamental frequency vector of the source speaker, selecting a STSS, or a series of STSS from the database for concatenation. The converted voiced POS is reconstructed using an adapted PSOLA approach at pitch marks of the source speaker in voiced regions and modulated by the amplitude of the incoming source speech waveform in order to maintain signal intensity.

### 3. SEGMENTATION AND MAPPING

To generate the set of units of the target speaker to concatenate, an incoming source signal is analyzed to detect and save the STSS for vowel reconstruction. Acoustic characteristics for each extracted segment of a vowel are thus analyzed, indexed, and mapped to the vowel-space.

According to the source filter model, a speech signal can be represented by

\[
s(n) = h_{VT}(n) \ast x_{EX}(n),
\]

where \( h_{VT}(n) \) describes the impulse response of the vocal tract, \( x_{EX}(n) \) the excitation signal coming from the lungs and \( s(n) \) is the speech waveform radiated at the lips [5]. If the excitation \( x_{EX}(n) \) is an impulse train, the spectral contour of a speech segment will approximate the frequency response of the excitation filter \( h_{VT}(n) \).

---

Figure 2: Block diagram for the applied voice conversion. The voiced parts-of-speech of an input signal from a source speaker \( \zeta \) are exchanged with a synthesized waveform using a series of acoustic units from a target speaker \( \xi \).

---

Figure 3: Block diagram of Fig. 2 represents the proposed voiced POS conversion system. During analysis, a target speaker’s voiced POS are each segmented into \( J \) segments \( s_j(n) \). The acoustic tracts. The main resonance frequencies of the vocal tract are called formant frequencies and mostly vary according to the perceived vowel quality.

The International Phonetic Association (IPA) issues a phonetic characterization of vowels according to the position of the tongue in the oral cavity at the time of utterance. Pitzinger [11] introduced a method which maps the acoustic characteristics of vowels to the cardinal vowel diagram of Fig. 1. The method uses the fundamental frequency \( f_0 \) and the first two formants \( F_1, F_2 \) as parameters such that

\[
h = 3.122 \log(f_0) - 8.841 \log(F_1) + 44.16 \quad (1)
\]

\[
b = 1.782 \log(F_1) - 8.617 \log(F_2) + 58.29, \quad (2)
\]

where \( h \) represents the perceived backness and \( b \) the perceived height of the tongue at the time of utterance according to the phone position on the IPA vowel diagram of Fig. 1. Backness is restricted between 0 and 12 and height between 0 and 10 to overcome the end-of-scale effect [11].

These coordinates will correspond to the Cartesian location of the VTSS in the vowel-space of Fig. 1, and are used as keys for selecting the units to concatenate in a database.

Synthesis of speech based on the concatenation of short-time signals in the time domain can be achieved with an acceptable fidelity in conjunction with a sinusoidal speech production model which conserves filter-source model theory [14, 15], and in sample-based granular synthesis, uses prerecorded segments with different voice qualities [12].

To exchange voiced parts-of-speech (POS) in the waveform, the signal is decomposed as in Fig. 2. An input speech signal from a source speaker \( x_\zeta(n) \) is divided into a signal containing voiced POS \( x_{V\zeta}(n) \) and a signal with other non-voiced (or non-tonal) POS \( x_{U\zeta}(n) \), such that

\[
x(n) = x_{V\zeta}(n) + x_{U\zeta}(n). \quad (3)
\]

The signal is analyzed to determine its vowel-quality every period. Sections of \( x_\zeta(n) \) are then re-synthesized using previously recorded grains (units) from a target speaker. The voiced region, \( x_{V\zeta}(n) \), is then exchanged with a synthesized voiced region of a target speaker \( x_{V\xi}(n) \), such to obtain the converted speech signal

\[
y(n) = x_{V\xi}(n) + x_{U\zeta}(n). \quad (4)
\]

The block diagram of Fig. 3 represents the proposed voiced POS conversion system. During analysis, a target speaker’s voiced POS are each segmented into \( J \) segments \( s_j(n) \). The acoustic

---

Figure 4: Voiced - Unvoiced segmentation.

---

3.1. Voiced and unvoiced detection

Voiced regions are analyzed using the fundamental frequency contour on the output of a YIN pitch tracker implemented as in [16] using a minimum pitch \( f_{p_{\text{min}}} = 50 \) Hz and a hop size of 256 samples.

Delimiters for voiced regions are set when the fundamental frequency surpasses 10% of the mean maximum frequency of a signal snippet. This is done to take in consideration the moment when a voiced tone starts to overcome the unvoiced POS [17]. For
our voice conversion application, regions shorter than 60 ms are ignored since the timbre of shorter sounds is not perceptually recognizable [18]. Fig. 4 shows the normalized \( f_0(n) \) and the voiced POS \( x_{V}(n) \) to be exchanged on the utterance /head/ of a male speaker from the Hillenbrand vowel database [19].

### 3.2. Segmentation and unit extraction

Phonetic segmentation is performed pitch-synchronously on the signal by fragmenting each voiced region \( x_{V}(n) \) of the target speakers into a STSS \( s_j(n) \) set for \( j \in [1, J] \).

The fundamental frequency vector \( f_0(n) \) at the output of the pitch tracker gives the cues for segmentation. Local maxima every pitch period length \( T_0(n) \) are identified as pitch marks on the voiced speech region. Maxima are searched for every \( \frac{1}{2}T_0(n) \) samples, so that no short-time speech segments overlap.

The first zero-crossing before such maxima is regarded as a delimiter of the STSS. Segments are indexed at consecutive delimiters, and form the units for concatenation during synthesis. Fig. 5 shows the segmentation.

### 3.3. Acoustic features

The acoustic features for each short-time speech segment \( s_j(n) \) are the first two formants \( (F_1, F_2) \) and the fundamental frequency \( f_0 \), which are used to calculate the perceived vowel-quality based on the backness and height coordinates on the vowel-space of Fig. 1.

After extracting the fundamental frequency and finding the delimiters of each STSS \( s_j(n) \), Linear Predictive Coding (LPC) [5, 20] is used to approximate the spectral envelope and the formant composition of the segment. The order of the LPC analysis filter is selected such that there is one pole for each kHz of the sampling frequency, in this case \( p = 44 \). The resulting filter coefficients are transformed to the frequency domain generating a spectral envelope approximation for the STSS.

A peak-picking technique is then used to obtain the frequency positions of the first two peaks in the envelope.

In Fig. 6 the spectral envelope of an /e/ show the formant composition of the segment. The peaks indicate the position of the formants in the frequency domain (the first three formants appear below 3 kHz).

### 3.4. Vowel mapping

A mapping technique is then used to transform the acoustical characteristics of the STSS to a position in the vowel-space. The amount of correlation which is introduced by the non-uniform Cartesian vowel coordinate system is settled by trans-
forming the vowel diagram’s trapezoidal form to a rectangular space using

\[ B = 10^b + 0.5h - 5 \]

\[ 0.5h + 7.25 \]

as in [21]. The position of \( s_j(n) \) on the vowel-space represents its perceived vowel quality, so that each extracted STSS represents a particular phone.

In order to reproduce non-isolated vowels, such as diphthongs and consonant - vowel transitions, the trajectories of the segments in the vowel-space should be considered [17]. Fig. 7 shows the perceived vowel height and backness for the STSS of two utterances of the English word /hawed/ from the Hillenbrand database [19]. The figure shows the transition between a starting STSS and an ending STSS in the (squared) vowel diagram from [21]. The perceived backness and height position of the segments also appear to change their velocity, transitioning slowly at the beginning and more abruptly towards the end of the voiced part.

### 3.5. Database construction

The database is based on the units required to synthesize the converted vowel waveform. In this sense, the window \( w(n) \) in Fig. 5 is used for saving units in the database. Each stored unit is situated between three consecutive delimiters for each extracted \( s_j(n) \). The Hann window \( w(n) \) is centered at the maxima of \( s_j(n) \) and expands over 1.5 times of the length of \( s_j(n) \). The window is then multiplied with a segment \( s_j^W(n) \), which is also centered at the peak of \( s_j(n) \) and is of the same length as \( w(n) \). Each segment \( w(n) - s_j^W(n) \) is saved in a dictionary along with its index \( j \), and backness \( B \) and height \( h \) positions from (2) and (1) respectively. The length of the segments is proportional to the period of the fundamental frequency as they were extracted.

The phonetic dictionary is then conformed of a series of segments \( s_j(n) \) extracted during analysis. Each series of grains, as in the vowelworm from Fig. 8, are saved. The keys for each extracted unit, correspond to the perceived backness from Eq. (6) and perceived height coordinates from Eq. (7). The trajectories of the diphones are also saved and a time stamp for the STSS is also added as a key. The units are normalized to its peak value and saved.

The target speaker’s database comprises the whole extracted segment grains. From Fig. 7, several interpretations can be extracted. The signal position on the vowel-space may describe the composition of a phoneme based solely on the use of steady-state vowel grains. The number of grains on each time position, the velocity and the start and ending positions should be considered to model the corresponding vowel. However, the pulses which correspond to steady state vowels could be modeled with a single grain as in [2].

The bi-dimensional matrix with the coordinates of the extracted segments for a speaker in the vowel dictionary is then

\[ \Gamma = [B, h], \]

where \([B, h]\) is a list of backness and height coordinates for every indexed segment in each vowel group, and \( v \) is the index for a particular vowel group. This will make up the keys for identification of a particular vowel segment \( s_j(n) \) in the dictionary of speaker-dependent vowel segments.

A vowel centroid is defined for each extracted voiced POS of the target speaker in the vowel-space such that

\[ \Gamma_v = \left[ \frac{1}{J} \sum_{j=1}^{J} B_j, \frac{1}{J} \sum_{j=1}^{J} h_j \right], \]

which will then form the bi-dimensional matrix of vowel centroids \( \Gamma \) for each speaker in the database. These centroids are used as keys for STSS retrieval.

### 4. SYNTHESIS AND VOWEL RECONSTRUCTION

Each STSS \( s_j(n) \) represents a particular phone whose length is only one period long and is intended to be used as a concatenation unit. However, PSOLA requires a greater synthesis window to re-generate speech without excessive artifacts [7]. The PSOLA synthesis window is thus set to 1.5 of the period length from the source speaker at each pitch mark.

Synthesis is performed by multiplying the speech waveform with a sequence of time-translated windows whose length is proportional to the local pitch period of the source speaker. The generated signal is analog to convolve an impulse train, whose impulses are located around glottal closure instants, with the steady-time response of the vocal tract at a particular utterance [13, 12].

#### 4.1. Unit selection

At each pitch mark \( k \), a STSS from the source speaker’s waveform is retrieved. One period of the signal is extracted as described in section 3 and the main two formants of extracted segment are calculated using the LPC approximation as in section 3.3. Backness and height measures for each segment are then calculated using Eqs. (6) and (1) respectively.

The resulting \([B, h]_k\) coordinates at the time instant \( k \) of the source speaker is used to map the corresponding vowel centroid in the vowel-space of the target speaker. The distance measure

\[ v = \min_{\Gamma} D ([B, h]_k, \Gamma), \]

is used, where \( \Gamma \) is the matrix of vowel centroids from section 3.5 of the target speaker and \( v \) is the resulting vowel group of the target
4.3. Speech signal reconstruction

Following Eq.4 the signal is reconstructed by fading the voiced parts \( \hat{x}_{V,ζ}(n) \) in the \( \hat{x}_{V,ζ}(n) \) signal after the voiced POS has been re-synthesized. This is done in an offline way. A Tukey window of length \( L \) is constructed, where \( L \) is the length of the voiced segment to fade. Following the, the slopes of the window are mirrored to reconstruct the speech waveform. Fig. 9 shows the procedure.

5. RESULTS AND EVALUATION

In this work, two sets of signals are considered for evaluation: Steady-state vowels and simple utterances from the Hillenbrand database [19].

To evaluate the algorithm on isolated vowels, several subjects where asked in a laboratory setting to record the same vowel order and retain intonation and intensity by following a prosodic stencil to avoid signal mismatches. The pitch lag of the source speaker is thus approximated to the length of each synthesis window, dropping the need for time stretching at the moment of concatenation. The recordings where composed of a succession of five primary vowels: /a/, /e/, /i/, /o/, /u/. The signals where thus solely composed of voiced regions of speech.

The converted speech waveform spectrum in Fig. 11c depicts that the resulting signal retains the pitch contour of the source
Acoustic units used for analysis are the STSS of a voiced phoneme. These units contain intrinsic vowel-quality and speaker specific features, and can be characterized as a phone.

The presented algorithm takes the prosodic properties (tonality, intensity, speech rate) of the speech signal from the source speaker and the filter characteristics of a target speaker are transferred to a source speaker. Subjective evaluation results for isolated vowels demonstrate that the speaker’s timbre is successfully transferred. Results of the synthesis of consonant-vowel transitions (Fig. 12) demonstrate that diphones (and other supra-segmental units) can be constructed using the phonetic transition visualization in the vowel-space.

The study of the transitions and characteristics of these short-time speech segments could give a better understanding of the phonological characteristics of speech and speakers, and generate a broader discussion in unit segmentation for speech analysis and synthesis. Study of the trajectories between phones in the vowel-space could be employed in real-time phonetic mapping of incoming speech, or to reduce unit inventories for speech conversion, for example.

In future work, speakers models for VC and TTS synthesis can be constructed based on the STSS positions in the vowel-space. Objective and perceptual measures for the presented speech conversion method should also be developed.

7. REFERENCES


Figure 11: Results for voice conversion for isolated vowels on continuous speech. The figure shows the spectrum of the signals for a source speaker (a), a target speaker (b) and the converted signal (c).

Figure 12: Results for voice conversion on consonant-vowel trajectories. The harmonic composition of the generated waveform corresponds to those of the target speaker.


