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. Phonolgy 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 synthesis-by-rule of the target speaker's vowel segments using the source speaker's tonality and intensity.

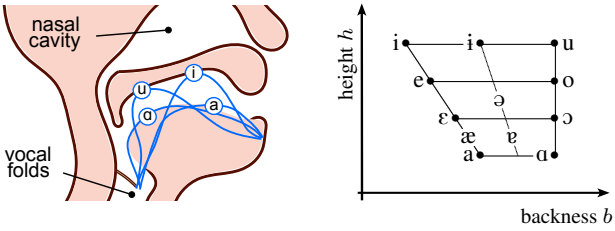


Figure 1: *Position of the tongue at the time of utterance and the cardinal vowel diagram. The x-axis corresponds to backness b and the y-axis to height h of the tongue at the time of utterance.*

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

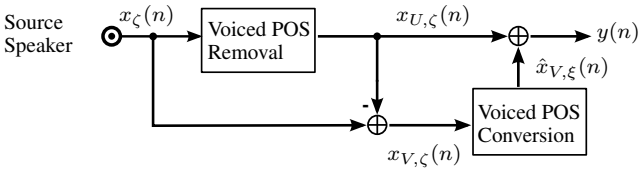


Figure 2: *Block diagram for the applied voice conversion. The voiced parts-of-speech of an input signal from a source speaker ζ are exchanged with a synthesized waveform using a series of acoustic units from a target speaker ξ.*

tract. 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. Pfitzinger [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 b represents the perceived *backness* and h 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 STSS in the vowel-space of Fig. 1, and are used as keys for selecting the units to concatenate out of a database.

Synthesis of speech based on the concatenation of short-time signals in the time domain can be achieved with an acceptable naturalness using PSOLA [12], where the signal to concatenate is overlapped at a specific position at each pitch mark k and added to the output signal [13]. PSOLA is used in several speech synthesizers 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

$$x(n)_\zeta = x_{V,\zeta}(n) + x_{U,\zeta}(n). \quad (3)$$

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 $\hat{x}_{V,\xi}(n)$, such to obtain the converted speech signal

$$y(n) = \hat{x}_{V,\xi}(n) + x_{U,\zeta}(n). \quad (4)$$

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 features of each STSS $s_j(n)$, the fundamental frequency f_{0_j}, and the first and second formants (F_{1j}, F_{2j}) 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

$$s(n) = h_{VT}(n) * x_{EX}(n), \quad (5)$$

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 $h_{VT}(n)$.

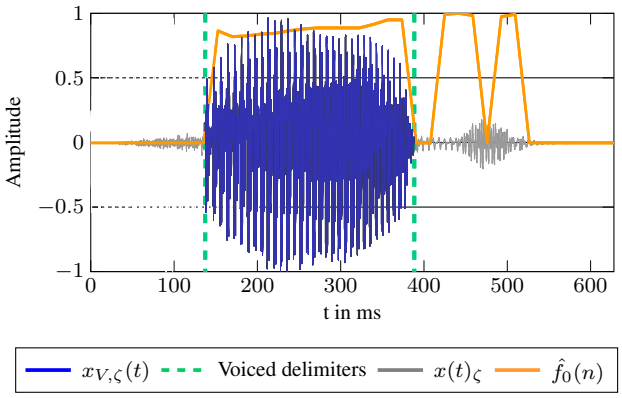


Figure 4: *Voiced - Unvoiced segmentation.*

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 $f_{0_{min}} = 50\,\text{Hz}$ 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

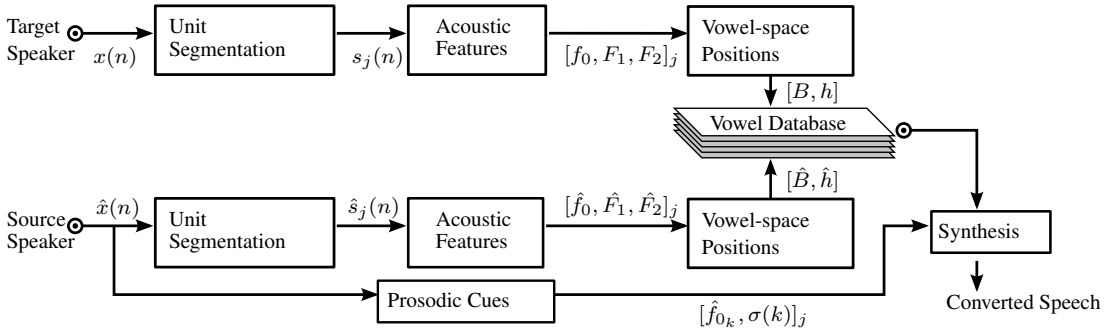


Figure 3: *Block diagram representation of the proposed vowel conversion system. Voiced speech segments of one period length are analyzed and exchanged with those of a target speaker.*

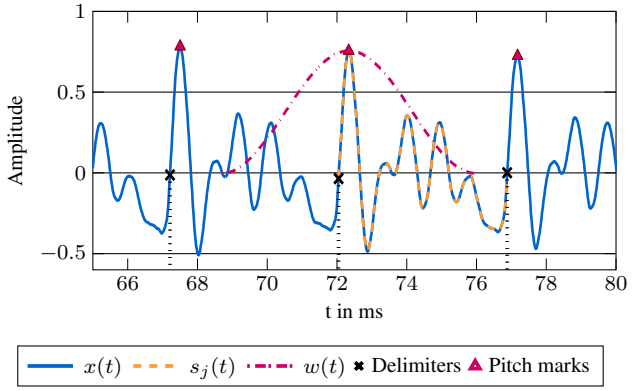


Figure 5: *Excerpt of a voiced segment (/a/). The dotted line shows the synthesis window $w(t)$ for samples in the database. Crosses represent the delimiters of the short-time speech segment $s_j(t)$. The window is centered at the local peak of each segment, i.e. the pitch marks.*

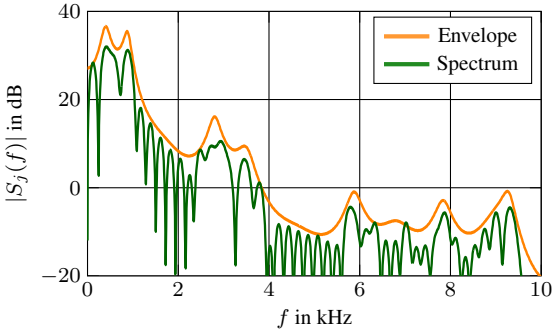


Figure 6: *LP spectral envelope and Fourier spectrum for a short-time speech segment of an /e/. The frequency positions of the envelope correspond with the formants.*

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 $\hat{f}_0(n)$ and the voiced POS $x_{V,\zeta}(n)$ 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 $\frac{3}{2}T_0(n)$ 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 $(F1, F2)$ 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.

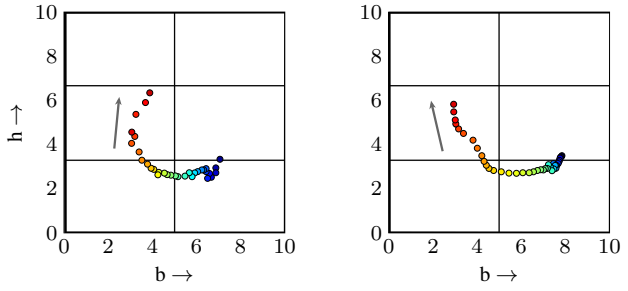
A peak-picking technique is then used to obtain the frequency positions of the first two peaks in the envelope.

In Fig. 6 the spectral envelope of an /e/ show the formant composition of the signal. 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-



(a) m09aw.wav (b) m01aw.wav

Figure 7: *Vowelworm [22] of the utterance /hawed/ from two different male speakers taken from the Hillenbrand database [19]. Each dot represents a short-time speech segment under analysis. The arrows indicate the approximate direction of the trajectories of the STSS in the vowel-space.*

forming the vowel diagram's trapezoidal form to a rectangular space using

$$B = 10\frac{b + 0.5h - 5}{0.5h + 7.25} \tag{6}$$

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 $s_j^D(n)$, which is also centered at the peak of $s_j(n)$ and is of the same length as $w(n)$. Each segment $w(n)\cdot s_j^D(n)$ 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 where 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.(**??**). 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

$$\boldsymbol{\gamma}_v = [\mathbf{B},\ \mathbf{h}]\,, \tag{7}$$

where $[\mathbf{B}, \mathbf{h}]$ 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

$$\boldsymbol{\Gamma}_v = \left[\frac{1}{J}\sum_{j=1}^{J} B_j, \frac{1}{J}\sum_{j=1}^{J} h_j\right], \tag{8}$$

which will then form the bi-dimensional matrix of vowel centroids $\boldsymbol{\Gamma}$ 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

$$v = \min_{v} \ D\left([B, h]_k, \boldsymbol{\Gamma}\right), \tag{9}$$

is used, where $\boldsymbol{\Gamma}$ is the matrix of vowel centroids from section 3.5 of the target speaker and v is the resulting vowel group of the target

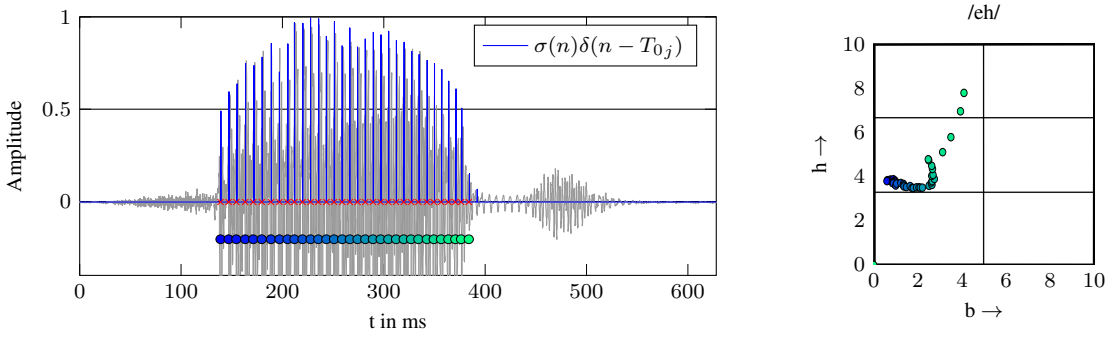


Figure 8: *Vowel segmentation and corresponding vowel-space trajectory. Each short-time speech segment (STSS) is pitch-synchronously extracted between the zero-crossings before each peak. Perceived tongue height and backness is approximated for each STSS, thus giving the positions of the segments in the vowel-space of the right. The generated impulse train (in blue) is used to give the pitch marks for re-synthesis. The impulse train is weighted with the source speaker's envelope for clarity reasons.*

speaker. The retrieved index corresponds to the vowel group of the target speaker closest to the uttered phone of the source speaker. The euclidean distance function

$$D(v)|_{[B,h]_k} = \sqrt{(B - \mathbf{\Gamma}_{(v,1)})^2 + (h - \mathbf{\Gamma}_{(v,2)})^2} \;\; \forall v, \qquad (10)$$

is applied to find the closest position in the $[\mathbf{B}, \mathbf{h}]$ vowel-space between two speakers.

A second search on the list of indexed STSS of the vowel group, $\mathbf{\Gamma}_v$ using (10) is also performed to find the index j of the STSS to concatenate. For consonant-vowel transitions the trajectories depicted in Fig. 7 are also considered in order to select the series of STSS to concatenate.

4.2. Vowel reconstruction

During voiced speech production, the excitation function $x(n)$ of (5) is represented as a sequence of impulses at the local pitch period $T_0(n)$ of the source speaker ζ. The fundamental frequency $f_0(n)$ is tracked at the source speaker as described in section 3.

Prosodic cues of the source speaker, such as intonation, intensity and vowel length are transfered to the converted signal by using the source speaker's fundamental frequency f_{0_k} and highest peak $\sigma(k)$ at each STSS to be replaced.

The synthesized waveform $\hat{x}_{V,\xi}(n)$ is the converted voiced POS and can be approximated to the synthesis model

$$\hat{x}_{V,\xi}(n) = \sigma(k) \sum_k w\left(n - kN_k\right) s_k^D(n), \qquad (11)$$

where $\sigma(k)$ is the time envelope of the source signal at the period maximum, $w(n-kN_k)$ is a variable Hann window of length N_k at each pitch event k of the source signal, and $s_k^D(n)$ is the unit segment from section 4.1 to concatenate. Discontinuities are avoided with the use of a hop size of $\frac{1}{4}N_k$.

The center of the window function is situated at the highest peak of the source's STSS, which can be approximated to its center of gravity [15].

4.3. Speech signal reconstruction

Following Eq.4 the signal is reconstructed by fading the voiced parts $\hat{x}_{V,\xi}(n)$ in the $\hat{x}_{U,\zeta}(n)$ 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. Followingly, the slopes of the window are mirrored to reconstruct the speech waveform. Fig. 9 shows the procedure.

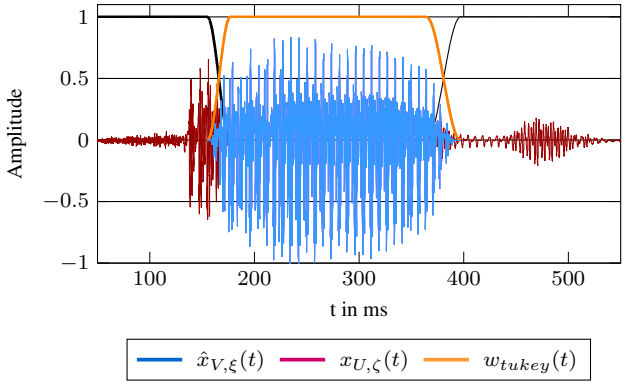


Figure 9: *Overlap add of the converted voiced part $\hat{x}_{V,\xi}(n)$ into the source signal $\hat{x}_{U,\zeta}(n)$.*

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

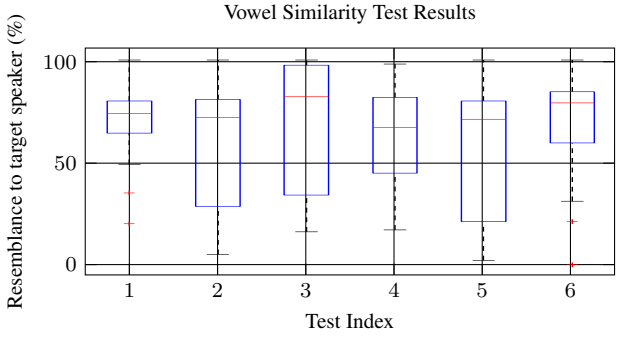


Figure 10: *Box-plot of the results for the evaluation test. The line inside each box indicates the mean of the decisions. Boxes closer to the top indicate that the identity of the speaker in the converted signal is subjectively closer to the target speaker.*

speaker of Fig. 11a while achieving the same energy for the harmonic composition of the target speaker of Fig. 11b. Vowel-quality and prosodic characteristics from the source speaker are thus transfered to a synthesized waveform with speaker-dependent characteristics of a target speaker.

These signals where also used to perform a subjective evaluation test on the capacity of the system to convert the identity of the source speaker to the target speaker. The evaluation is based on the ABX test [23, 24]. The test measures the degree of match between a source (A), a target (B), and a converted (X) signal. The tester should decide whether the converted sample (X) is perceptually closer to the identity of the source (A) or the target (B) speaker. A group of twenty five independent testers with basic to professional musical knowledge and familiar with signal processing where instructed to take the test. The test consisted on 6 different sets of a source, a target and a converted signal of the vowels between two speakers. The spectrum of the signals in the first set is shown in Fig. 11. The box-plot of Fig. 10 presents the outcome of the results. It can be noticed that the mean of the decisions is closer to the target speaker for all the signal sets.

For evaluation on consonant-vowel transitions, several samples of male and female speakers from the Hillenbrand database where taken in consideration. Figure 12 shows the spectrum of a source and a target speaker, as well as the converted speech sample. The converted signal contains re-synthesized voiced POS with the vocal tract information (timbre) of a target speaker and the unvoiced POS of a source speaker (as well as voiced regions shorter than 60 ms). The spectrum shows the same case as with isolated vowels, where the harmonic composition of the voiced regions of converted speech resemble those of the target speaker, while preserving prosodic features like phoneme duration and tonality of the source speaker.

6. CONCLUSIONS

An automatic vowel conversion system is presented which uses a short analysis and synthesis window for speech morphing and a vowel quality mapping method for period-length segments, or STSS.

A low dimensional acoustic parameter for voiced phonemes (formants, vowel-quality) is used to recognize the five vowels /a/, /e/, /i/, /o/ and /u/ according to the position in the vowel-space. The 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 transfered to a *source* speaker. Subjective evaluation results for isolated vowels demonstrate that the speaker's timbre is successfully transfered. 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 visualized 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

[1] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using speaker-dependent conditional restricted boltzmann machine," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–12, 2015.

[2] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 67–79, 2007.

[3] G. Hickok, "The architecture of speech production and the role of the phoneme in speech processing," *Language, Cognition and Neuroscience*, vol. 29, no. 1, pp. 2–20, 2014.

[4] H. Kawahara, H. Banno, T. Irino, and P. Zolfaghari, "Algorithm amalgam: morphing waveform based methods, sinusoidal models and STRAIGHT," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2004*, vol. 1, pp. I–13.

[5] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech Synthesis Based on Hidden Markov Models," *Proc. of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[6] H Hon, A. Acero, X. Huang, J. Liu, and M. Plumpe, "Automatic generation of synthesis units for trainable text-to-speech systems," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 1998*, vol. 1, pp. 293–296.

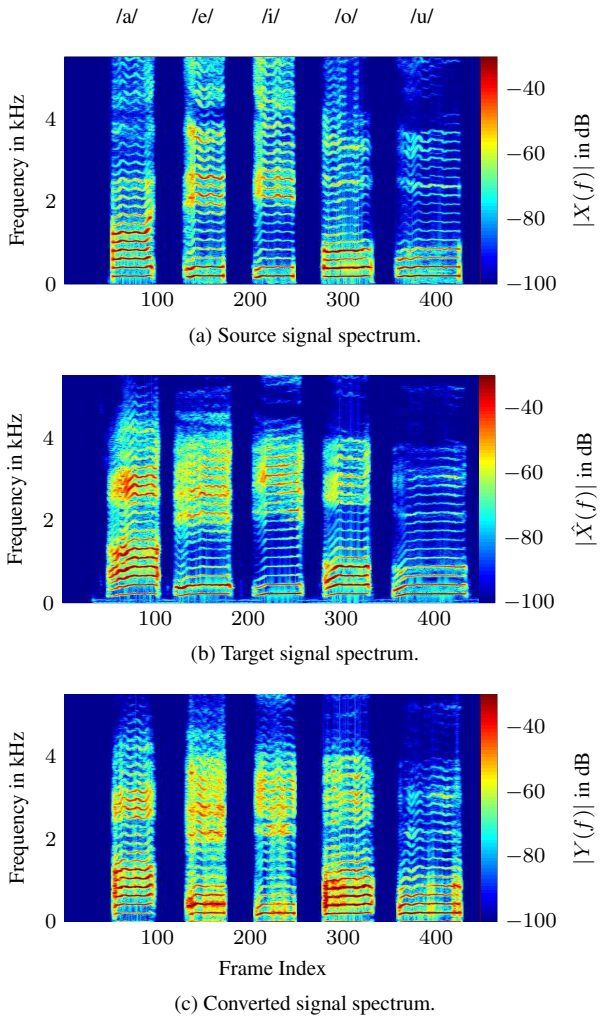
[7] R. Donovan and P. Woodland, "A hidden markov-model-based trainable speech synthesizer," *Computer speech & language*, vol. 13, no. 3, pp. 223–241, 1999.

[8] T.F. Cleveland, "Acoustic properties of voice timbre types and their influence on voice classification," *The Journal of the Acoustical Society of America*, vol. 61, no. 6, pp. 1622–1629, 1977.

[9] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[10] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 1986*, vol. 11, pp. 2015–2018.

[11] H.R. Pfitzinger, "Acoustic correlates of the IPA vowel diagram," *Proc. of the XVth Int. Congress of Phonetic Sciences*. Citeseer, 2003, vol. 2, pp. 1441–1444.

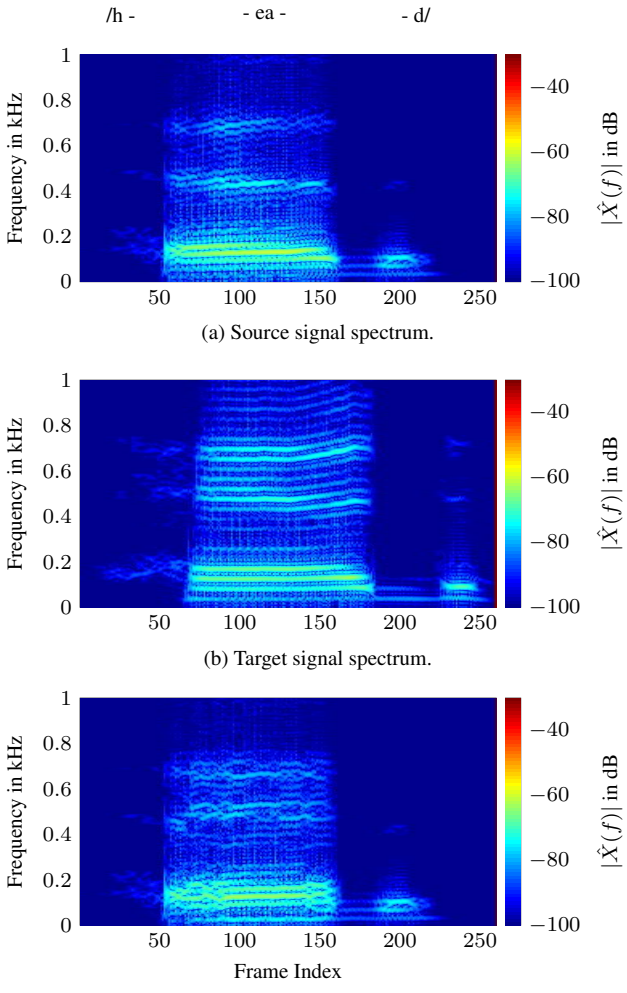


(a) Source signal spectrum.

(b) Target signal spectrum.

(c) Converted signal spectrum.

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).*



(a) Source signal spectrum.

(b) Target signal spectrum.

(c) Converted signal spectrum.

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

[12] X. Sun, "Voice quality conversion in TD-PSOLA speech synthesis," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2000*, vol. 2, pp. II953–II956.

[13] C. Hamon, E. Mouline, and F. Charpentier, "A diphone synthesis system based on time-domain prosodic modifications of speech," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 1989*, pp. 238–241.

[14] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter, "TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 1998.*, vol. 1, pp. 273–276.

[15] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Tran. on Speech and Audio Processing,* , vol. 9, no. 1, pp. 21–29, 2001.

[16] U. Zölzer, *DAFX digital audio effects*, John Wiley & Sons, 2011.

[17] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PloS one*, vol. 8, no. 4, 2013.

[18] S.Z.K. Khine, T.L. Nwe, and H. Li, "Exploring perceptual based timbre feature for singer identification," *Computer Music Modeling and Retrieval*, Springer, pp. 159–171, 2008.

[19] J. Hillenbrand, L.A. Getty, M.J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *The Journal of the Acoustical society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[20] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Tran. on Acoustics, Speech and Signal Processing*, vol. 22, no. 2, pp. 135–141, 1974.

[21] H.R. Pfitzinger, "Towards functional modelling of relationships between the acoustics and perception of vowels," *ZAS papers in Linguistics*, vol. 40, pp. 133–144, 2005.

[22] H. Frostel, A. Arzt, and G. Widmer, "The vowel worm: Real-time mapping and visualisation of sung vowels in music," *Proc. of the 8th Sound and Music Computing Conference*, pp. 214–219, 2011.

[23] J. Kreiman and G. Papcun, "Comparing discrimination and recognition of unfamiliar voices," *Speech Communication*, vol. 10, no. 3, pp. 265–275, 1991.

[24] T. Ganchev, A. Lazaridis, I. Mporas, and N. Fakotakis, "Performance evaluation for voice conversion systems," *Text, Speech and Dialogue*. Springer, pp. 317–324, 2008.