Voicing Level Control with Application in Voice Conversion

Jani Nurminen¹, Jilei Tian², and Victor Popa²

¹ Nokia Technology Platforms, Tampere, Finland
² Interaction Core Technology Center, Nokia Research Center, Tampere, Finland

{janik.nurminen, jilei.tian, ext-victor.popa}@nokia.com

Abstract

Speech processing related changes in the speech spectra may often lead to unwanted changes in the effective degree of voicing, which in turn may degrade the speech quality. This phenomenon is studied more closely in this paper, first on a theoretical level and then in the context of voice conversion. Moreover, a simple but efficient approach for avoiding the unwanted changes in the effective level of voicing is proposed. The usefulness of the proposed voicing level control is demonstrated in a practical voice conversion system. The compensation of the changes in the degree of voicing is found to reduce the average level of noise in the output and to enhance the perceptual speech quality.

Index Terms: voicing, spectral modification, voice conversion

1. Introduction

The concept of voicing is one of the fundamental concepts in speech processing. Roughly speaking, from the speech production point of view, voiced sounds are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in oscillation. Unvoiced sounds, on the other hand, are produced without the oscillation of the vocal cords. In human speech, vowels are usually considered voiced while consonants can be voiced or unvoiced. Voicing related information can be utilized in many ways in speech processing.

In some very simplified speech models, such as in the basic source-filter model used in the well-known linear prediction coding (LPC) based LPC-10 vocoder [1], all segments of speech are regarded as either fully voiced or fully unvoiced. The voiced excitation is modeled as periodic pulses while the unvoiced excitation is represented using random noise. Even though this approach is quite well in line with the simplified view on human speech production, it has been found inadequate for producing high quality speech. A natural way for improving the performance is to allow different degrees of voicing in the excitation modeling. There are many successful speech models that utilize this idea in different ways. For example, the model used in waveform interpolation (WI) speech coding [2] allows voiced and unvoiced contributions to co-exist everywhere in the spectrum in the form of slowly and rapidly evolving waveforms, while the model used in multi-band excitation (MBE) speech coding [3] separates the speech spectrum into multiple frequency bands that can each be either voiced or unvoiced.

In many speech processing applications, the speech signal is modified in a manner that causes changes in the spectrum. Typical examples of such cases include the coarse quantization of the linear prediction coefficients in very low bit rate speech coding, the spectral smoothing at concatenation boundaries in concatenative text-to-speech (TTS) synthesis, and modification of the spectra in voice conversion where the aim is to convert the speech of a source speaker to sound as if uttered by a second speaker referred to as the target speaker. This kind of modifications performed on speech signals or their spectra may also cause unwanted changes in the effective degree of voicing if there is no explicit control on voicing. The changes in voicing, in turn, may degrade the perceptual quality of the processed speech.

In this paper, we study the problem of unwanted changes in the degree of voicing and propose explicit voicing control to tackle it. The proposed approach for voicing control is implemented and experimented with in a slightly modified version of the voice conversion system presented earlier in [4]. The voicing control is found to offer more natural and stable voicing levels and a clear improvement in speech quality.

This paper is organized as follows. In Section 2, we show that spectral modifications may lead to unwanted changes in the effective level of voicing. A simple but efficient solution to the problem is introduced in Section 3. The changes in voicing and the proposed solution are further discussed in Section 4 from the viewpoint of voice conversion. Finally, Section 5 concludes the paper.

2. Unwanted changes in voicing

Many speech processing techniques utilize linear prediction (LP). For this reason, and to make the discussions easy to follow, it is assumed in this paper that the spectral envelope of the vocal tract contribution is modeled using linear prediction. Moreover, we assume that the excitation is modeled using a sinusoidal model proposed e.g. in [5] and [6]. Nevertheless, it should be noted that this approach was chosen only for convenience and that the voicing related phenomenon discussed in this paper is not only present in this particular speech model or in voice conversion.

2.1. Definitions

In the well-known linear prediction approach, previous samples are used in forming a prediction for a new sample. The output of the LP analysis filter can be computed as

\[ r(t) = s(t) - \sum_{j=1}^{K} a_j s(t-j), \]

where \( s(t) \) denotes the discrete speech signal value at time \( t \), \( K \) is the filter order, and \( r(t) \) denotes the residual signal that cannot be predicted. The linear prediction coefficients \( \{a_j\} \) are generally estimated using the autocorrelation method or the covariance method, with the former being more popular due to the ensured filter stability.

The LP coefficients can usually model the vocal tract contribution reasonably well. Following the well-known source-filter interpretation, the remaining residual \( r(t) \) can be
regarded as the excitation signal. The excitation signal can be modeled in many ways. Here, we assume for simplicity and notational convenience that the excitation is modeled in a frame-wise manner as a sum of sinusoids,

\[
    r(n) = \sum_{m=1}^{M} A_m \cos(n\omega_m + \theta_m),
\]

where \(A_m\) and \(\theta_m\) represent the amplitude and the phase of each sine-wave component associated with the frequency track \(\omega_m\) and \(M\) denotes the total number of sine-wave components.

The voicing can be taken into account in different ways. One alternative would be to model only the voiced contribution using Eq. (2) and the unvoiced contribution could be modeled separately as a spectrally shaped noise. In this paper, we choose to also model the unvoiced portion using sinusoids, leading to the model

\[
    r(n) = \sum_{m=1}^{M} A_m [v_m \cos(n\omega_m + \theta_m) + (1-v_m) \cos(n\omega_m + \theta'_m)],
\]

where \(v_m\) is the degree of voicing for the \(m\)th sinusoidal component ranging from 0 to 1, while \(\theta_m\) and \(\theta'_m\) denote the phase of the \(m\)th voiced and unvoiced sine-wave component, respectively. A reasonably accurate approximation, from the perceptual point of view, can be obtained using linearly evolving voiced phases and random unvoiced phases.

2.2. Changes in voicing

To illustrate the unwanted changes in voicing, let us assume that the original LP coefficients are modified from \(\{a_i\}\) to \(\{a'_i\}\) as a result of some speech processing technique. The modification could happen e.g. due to very coarse quantization in very low bit rate speech coding, due to spectral smoothing in concatenative TTS or due to a voice conversion related transformation. As a result of this modification, the spectral envelope changes accordingly. Assuming that the filter remains stable (that can be guaranteed e.g. by performing the modification in the line spectral frequency domain), the old and the new spectral envelopes can be directly computed based on the LP synthesis filter using

\[
    H(e^{i\omega}) = \frac{1}{1-\sum_{j=1}^{K} a_j e^{-i\omega}},
\]

and by using the same equation with the modified coefficients \(\{a'_i\}\) to obtain \(H'(e^{i\omega})\).

The effect that the spectral modification has on voicing can be studied by measuring the energies of the voiced and unvoiced contributions in the spectrum before and after the modification. The average energy of the voiced part for a single frame, denoted as \(E_V\), can be estimated by sampling the spectrum at the frequencies of the sinusoids, \(\omega_m\), as

\[
    E_V = \sum_{m=1}^{M} \left| H(e^{i\omega_m})v_m A_m \right|^2.
\]

Similarly, the energy of the unvoiced contribution, \(E_U\), can be computed as

\[
    E_U = \sum_{m=1}^{M} \left| H(e^{i\omega_m})(1-v_m) A_m \right|^2.
\]

The corresponding energies after the spectral modifications, \(E'_V\) and \(E'_U\), can be obtained using similar calculations as in Eq. (5) and Eq. (6) but by substituting \(H(e^{i\omega})\) with \(H'(e^{i\omega})\) to take into account the changes in the LP coefficients. It should also be noted that if the spectral modifications would cause changes in other parameters than the LP coefficients, these changes should also be taken into account when computing \(E'_V\) and \(E'_U\).

It is usual in speech processing systems to carefully control the behavior of the overall energy but it is not common to explicitly control the relative contributions of the voiced and unvoiced components to the overall energy. However, if there is no explicit control on voicing, the spectral modifications often change the perceived level of voicing in a clearly audible way. This is caused by the fact that the relative contribution of the voiced (or the unvoiced) component to the overall energy is often changed due to the spectral modification, i.e.

\[
    \frac{E'_V}{E'_V+E'_U} \neq \frac{E_V}{E_V+E_U}.
\]

The perceptual effect of the unwanted changes in voicing can in practice be observed as audible changes in the amplitude of the spectrally shaped noise generated to model the noise-like unvoiced contribution. This effect is discussed more closely and demonstrated using a practical example and experimental results in Section 4.

3. Voicing control

The unwanted changes in voicing can be corrected by controlling the voicing in an explicit way. The detailed implementation of the voicing control depends on the speech model used in the target application. In addition, even with a fixed speech model, there are different alternatives regarding the implementation.

In the case of the speech model discussed in Section 2, one possible solution could be to establish a frequency-dependent function for modifying the degree of voicing for the different sinusoids. For example, in the simplest case,

\[
    v'_m = f(v_m, \omega_m, E_V, E_U).
\]

The exact function can be designed in many ways and the parameters used in defining the modified degree of voicing, \(v'_m\), could also be different than the ones given in Eq. (8).

Nevertheless, the aim is to modify the voicing of the sinusoids in such a manner that if computations similar to Eq. (5) and Eq. (6) would be applied again for calculating the energies after the voicing control, \(E'_V\) and \(E'_U\), we would now have

\[
    \frac{E'_V}{E'_V+E'_U} = \frac{E_V}{E_V+E_U}.
\]

Alternatively, it may be desired to only go towards this goal without fully satisfying it, or the target level of voicing might be decided using other techniques. E.g. in voice conversion, there could be a separate conversion model for finding out the target levels for the relative contributions of the voiced and unvoiced components based on the non-converted voicing values and possibly some other parameters, with the aim of modeling the speaker-dependencies in voicing.

The frequency-dependent operation sketched above can be used, for example, for focusing the increase in voicing more to low frequencies and/or the decrease in voicing more...
to high frequencies, which may be perceptually justified. However, a much simpler but still effective solution can be obtained by treating all sinusoids in the same manner. It is easy to see that the objective can be approximately achieved e.g. using the following simplified function,

\[ v'_a = f_a(v_a, E_y, E'_y) = \min \left\{ v_a, \sqrt{\frac{E_y}{E'_y}} \right\}. \]  (10)

In cases where \( E'_y = 0 \), the voicing can be left unmodified.

Assuming that there is also a mechanism for ensuring that the overall energy stays unchanged, and that the voicing values \( v_a \) are continuous values in the range from 0 to 1, the simplified solution presented in Eq. (10) effectively controls the level of voicing. If the voicing decisions are hard as e.g. in the MBE model, i.e. \( v_a \) is always either 0 or 1, the best solution would be to change the voicing values of some sinusoids to approximately satisfy the condition in Eq. (9). Similar approach but with continuous voicing values could be used to complement the simplified solution in Eq. (10) to fully satisfy Eq. (9). Another solution could be obtained by also modifying the amplitudes of the sinusoids in addition or instead of modifying the degree of voicing of the sinusoids.

4. Practical application: voice conversion

In this section, we demonstrate the usefulness of the proposed voicing control in a voice conversion related application. The voice conversion system used in these experiments has been originally presented in [4]. The first part of this section gives a brief overview of the system while the second part discusses the practical experiments related to the voice control.

4.1. Voice conversion system

The voice conversion system originally presented in [4] is based on a parametric speech model. The model is also applicable to speech coding [7], and it is essentially similar to the one presented in Section 2, except for the fact that the sinusoids are always harmonically related, i.e. the frequencies of the sinusoids are integer multiples of the fundamental frequency \( \omega_b \). During voiced speech, \( \omega_b \) corresponds directly to the pitch associated with the analysis frame. During unvoiced speech, however, there is no physically meaningful pitch available, and we use a fixed value for \( \omega_b \).

The conversion of the speech parameters is generally handled one vector at a time, using a Gaussian mixture modeling (GMM) based approach. Let \( x \) and \( y \) denote parameter vectors associated with the source and the target speakers, respectively. For the training of a conversion model, combined source-target feature vectors are generated by joining aligned source and target vectors, denoted as \( \mathbf{x} = [\mathbf{x}^{\text{S}}, \mathbf{x}^{\text{T}}]' \), that are used for training a conversation model. In the training, we have used the popular approach proposed in [8] that makes use of the aligned data \( \mathbf{z} \) to estimate the GMM parameters \( (\alpha, \mu, \Sigma) \) of the joint density \( p(\mathbf{z}) \). This is accomplished iteratively through the well-known Expectation-Maximization (EM) algorithm [9]. The actual conversion follows a scheme where the trained GMM parameterizes a mapping function that minimizes the mean squared error (MSE) between the converted source and target vectors [8],

\[ F(x) = E(y | x) = \sum_{k=1}^{L} \alpha_{k} \cdot N(x; \mu_{k}^{y}, \Sigma_{k}^{y}) \] (11)

where

\[ p_{k}(x) = \frac{\alpha_{k} \cdot N(x; \mu_{k}^{y}, \Sigma_{k}^{y})}{\sum_{j=1}^{K} \alpha_{j} \cdot N(x; \mu_{j}^{y}, \Sigma_{j}^{y})}. \] (12)

The GMM parameter \( \alpha_{k} \) is the prior probability for the \( k \)th component, \( L \) denotes the number of mixtures, and \( N(x; \mu, \Sigma) \) is the Gaussian distribution with mean \( \mu \) and covariance \( \Sigma \). Moreover, \( \alpha_{k} \) are non-negative and sum up to one. The covariance matrix of the 4th mixture is constructed as

\[ \Sigma_{4} = \begin{bmatrix} \Sigma_{4}^{xx} & \Sigma_{4}^{xy} \\ \Sigma_{4}^{yx} & \Sigma_{4}^{yy} \end{bmatrix}. \] (13)

Similarly,

\[ \mu_{k} = \begin{bmatrix} \mu_{k}^{x} \\ \mu_{k}^{y} \end{bmatrix}. \] (14)

represents the mean vector of the 4th Gaussian mixture.

The speech parameters are converted using the above GMM based approach together with the data clustering and mode selection technique presented in [10]. The exception to this is the pitch parameter that is converted using the prosody conversion technique introduced in [11]. The voicing values are kept unchanged in the conversion since we have not found clear speaker-dependencies in the voicing values. The changes in the pitch are taken into account by modifying the total number of the sinusoids \( M \) and the frequencies \( \omega_{b} \), accordingly, and by re-sampling the underlying spectral information at these new frequencies using interpolation.

4.2. Effect of voicing control

In voice conversion, the spectral changes are coming from multiple sources because many, if not all, parameters are converted. However, the voicing control can still be implemented as proposed earlier in this paper, provided that the changes in all parameters are taken into account when applying the equations proposed in Section 2 and Section 3. The voicing control can operate directly on the voicing values without changing any other parameter values.

A practical example case demonstrating the need for controlling the voicing is given in Figure 1. The figure depicts the overall level of voicing, calculated as the ratio between the energy of the voiced contribution and the total energy, \( E_{v} / (E_{v} + E_{n}) \), for each 10-ms frame of an example sentence before and after voice conversion. As can be seen from the figure, the effective voicing level has clearly changed in the conversion even though the parameter values related to the degree of voicing have not been converted at all. The figure also shows that the effective level of voicing is often decreased in the conversion, leading to a higher contribution of the noise-like excitation that can be perceptually observed as increased noise. Moreover, since the difference in the level of voicing before and after the conversion is not constant, the increase in noise is also non-constant, leading to a pumping-like perceptual effect.

The unwanted changes in voicing were also studied using a larger test set consisting of 42 sentences. The overall level of voicing was measured before and after voice conversion for all the 10-ms frames of these sentences. In 47.9% of the frames, the voice conversion decreased the level of voicing, while an increase in the level of voicing occurred in 26.4% of the frames. In the rest of the frames, the level of voicing did not change due to the fact that the whole spectrum was
considered either fully voiced or fully unvoiced both before and after conversion. The average level of voicing in the whole training set including all the frames was decreased due to the conversion by about 2.8%. These experimental results on a larger test set support the findings that were observed in the sentence illustrated in Figure 1: the voice conversion system on average decreases the level of voicing, leading to an increased level of noise in the output.

The proposed scheme for voicing control can efficiently correct the level of voicing in such a manner that the voicing change caused by the voice conversion system is always fully compensated. The perceptual effect of having the voicing control available was informally evaluated using four expert listeners. The listeners heard converted speech samples with and without voicing control and were asked to evaluate the quality differences between the two samples. Repeated listening was possible without any restrictions. The language used in the samples was English and the voice conversion was performed between genders.

The speech quality in the samples produced using voicing control was always observed better or equal to the quality of the conventional samples. The voicing control was found to remove part of the noise generated during the conversion. This is quite natural since the voicing control slightly reduces the contribution of the noise-like component, and it makes the behavior of that component more stable and ensures that it is better in line with the expected degree of voicing.

5. Conclusions
In this paper, we have discussed the effect that spectral modifications have on the degree of voicing. We have shown that modifications in the spectra may cause unwanted changes in voicing. Moreover, we have proposed a simple but effective scheme for controlling the voicing level. The usefulness of this approach was demonstrated in a voice conversion system. The inclusion of the proposed voicing control mechanism was found to reduce the level of noise in the output signal and to offer improvements in perceptual speech quality.

6. Acknowledgements
This work has partially been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech-to-Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

7. References