



Adding Glottal Source Information to Intra-lingual Voice Conversion

Javier Pérez, Antonio Bonafronte

Universitat Politècnica de Catalunya, Barcelona, Spain

Abstract

This paper studies the inclusion of glottal source characteristics in voice conversion (VC) systems. We use source/filter decomposition to parametrize the vocal tract using LSF, the glottal source using the LF model, and the aspiration noise using amplitude-modulated high-pass filtered AWGN noise. To evaluate the impact of this new parametrization in VC, we use a reference conversion system that estimates a linear transformation function using a joint target/source model obtained with CART and GMM. The reference system is based on the LPC model, uses LSF to represent the vocal tract and a selection technique for the residual. We use the reference algorithm to build a VC system for each of the three parameter sets. We compared both parametrizations in the framework of an intra-lingual voice conversion task in Spanish. The results show that the new source/filter representation clearly improves the overall performance, both in terms of speaker identity transformation and voice quality.

Index Terms: glottal modeling, voice conversion, speech analysis, speech synthesis.

1. Introduction

The aim of this paper was to investigate whether existing voice conversion (VC) methods would benefit from a more accurate speech production representation as initially proposed by Fant [1]: the speech $S(z)$ is produced when the waveform glottal source $U_g(z)$ excites the vocal tract $V(z)$, and is radiated by the lips $L(z)$ (approximated as a first-order differentiator). Modeling $V(z)$ as an all-pole filter and using linearity, we re-order the filters as:

$$S(z) = U_g(z)V(z)L(z) = G(z) \frac{1}{1 - \sum_{k=1}^N a_k z^{-k}}, \quad (1)$$

where we work with the derivative glottal volume-velocity waveform $G(z) = U_g(z) \cdot L(z)$ and a_k are the vocal-tract filter coefficients. In recent years there have been several methods aiming at the automatic estimation of the glottal source and the vocal tract. Most efforts focus into the independent estimation of the vocal tract, and then obtain the glottal waveform by inverse filtering the speech signal [2]. These methods often need to work using only speech segments corresponding to the glottis closed-phase, thus resulting in inaccurate estimations since it can often be very short (or non-existing). Hence, recent research focuses on the joint estimation of both the voice source and the vocal tract [3].

Recently we proposed a method belonging to this second category based on convex optimization techniques [4]. Our intention is to evaluate its performance in an intra-lingual voice conversion task using a reference VC system developed in our group for the TC-STAR European project [5]. This reference system is based on CART and GMM, and uses acoustic (LP) and phonetic characteristics. Although newer VC

paradigms [6] have been proposed which improve the conversion performance, we are here only interested in evaluating the impact of the signal parametrization. We can also expect newer methods to benefit from a better parametrization in a similar manner.

We start the paper with an explanation of the algorithm used for source-filter decomposition, in Section 2. Then we proceed to explain in Section 3 the reference voice conversion algorithm used in this work, and how the new parametrization is included. The subjective evaluation that we performed and its results are explained in Section 4. Finally, we end the paper with the conclusions and some directions for future work in Section 5.

2. Source-filter parametrization

In order to obtain the parameters of the voice source and the vocal tract, we first use our previously reported algorithm [4] to perform the source-filter decomposition, and then we proceed to the parametrization. For the sake of simplicity, we will only review here the most relevant parts of the algorithm, for the details please refer to the aforementioned reference.

2.1. Vocal tract

From eq. (1), we see that given a set of $N + 1$ filter coefficients (N for the vocal tract, and 1 for the tilt coefficient μ , see below), the speech signal $s(n)$ can be inverse-filtered to obtain an approximation of the glottal waveform:

$$g_{if}(n) = s(n) - \sum_{k=1}^{N+1} a_k s(n-k). \quad (2)$$

We use the KLGLOTT88 [7] model as an initial parametrization of the estimated glottal waveform (2), which is formulated as a Rosenberg-Klatt waveform:

$$g_{rk}(n) = \begin{cases} bn(2n_c - 3n) & , 0 \leq n < O_q T_0 \\ 0 & , O_q T_0 \leq n < T_0. \end{cases} \quad (3)$$

followed by a first-order low-pass filter controlling the smoothness of the glottis closure (i.e., spectrum tilt $TL(z) = \frac{1}{1-\mu z^{-1}}$). O_q is the duration of the open phase (%), T_0 is the glottal cycle length, b controls the amplitude, and $n_c = T_0 O_q$ is the glottal closure instant.

Our goal is to obtain the parameters b and a_k that minimize the norm of the parametrization error, when using the KLGLOTT88 model (3) to approximate the inverse-filtered waveform (2). To obtain smoother estimates and reduced period-to-period variability, we use segments of 3 consecutive glottal

periods:

$$e(n) = g_{rk}(n) - g_{if}(n) \quad (4)$$

$$= \begin{cases} b C_1(n) + \sum_k a_k s(n-k) - s(n) & n \in OP_1 \\ \sum_k a_k s(n-k) - s(n) & n \in CP_1 \\ b C_2(n) + \sum_k a_k s(n-k) - s(n) & n \in OP_2 \\ \sum_k a_k s(n-k) - s(n) & n \in CP_2 \\ b C_3(n) + \sum_k a_k s(n-k) - s(n) & n \in OP_3 \\ \sum_k a_k s(n-k) - s(n) & n \in CP_3 \end{cases}$$

where again we have simplified the notation using $C_i(n) = n(2n_c^i - 3n)$, and OP_i and CP_i are the open and closed phases for glottal cycle i inside the analysis frame; the summatories range from $k = 1$ to $k = N + 1$. The frames are updated on a period-by-period basis, and the results combined using a overlap-and-add procedure.

Since the error is linear w.r.t. our unknown variables (b and a_k), we can rewrite eq. (4) in matrix form: $\mathbf{e} = \mathbf{F}\mathbf{x} - \mathbf{y}$. The vector $\mathbf{x} = [b \ a_1 \ \dots \ a_{N+1}]^T$ contains the variables to be estimated, $\mathbf{y} = [s(1) \ \dots \ s(P)]^T$ consists of known speech sample and the matrix \mathbf{F} contains $C_i(n)$ in the first column and speech samples in the rest. Minimizing the L_2 norm of the error is a convex optimization problem, and thus guaranteed to result in the optimal global minimum [4]. As a result, we obtain an optimal set of filter coefficients a_k and the amplitude b of the glottal waveform g_{rk} parameters. We use the standard approach of representing the vocal tract filter using LSF due to its better interpolation properties, so the vector of parameters representing the vocal tract is $\theta_{vt} = [l_{s1} \ \dots \ l_{sN}]^t$.

2.2. Glottal source

The glottal waveform obtained by inverse-filtering (2), can be better approximated using the more flexible LF model [8]:

$$g_{if}(t) = \begin{cases} E_0 e^{\alpha t} \sin(w_g t) & , 0 \leq t \leq t_e, \\ -\frac{E_e}{\epsilon t_a} [e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}] & , t_e < t \leq t_c, \\ 0 & , t_c < t \leq T_0, \end{cases} \quad (5)$$

We estimate the LF parameters t_e , t_a , t_p and E_e for each pitch period by minimizing the L_2 norm between (2) and (5) using a non-linear, least-squares minimization procedure. t_o and t_c are initially set to 0 and $T_0 = 1/F_0$ and are not modified. Convolution with a length 7 Blackman window is used to low-pass filter the pulses and reduce the noise before error computation. The initialization step is crucial to obtain good estimates, so instead of relying on error-prone direct estimation, we map g_{kl} onto the LF space [4].

The glottal waveform is then modeled using the equivalent normalized LF parameters $R_a = t_a/T_0$, $R_k = t_e/t_p - 1$ and $R_0 = t_e \cdot T_0$, since they are better suited to prosody modifications as we will see later on. The vector of parameters representing the glottal source is $\theta_{lf} = [R_a \ R_k \ R_0 \ E_e]^t$.

2.3. Aspiration noise

At this point we could already synthesize speech using the estimated vocal tract and LF parameters in eq. 1. However, since we are not including any of the original aspiration noise, the resulting speech would sound quite unnatural, typical of a vocoder-like scheme. We will recover the aspiration noise from

the parametrization error, and use amplitude-modulated Gaussian noise to represent it. This follows from the results of turbulent noise theory [9], where the aspiration noise is found to mainly consist of two components: a constant leakage during the whole glottal cycle and a primary burst of noise occurring at the beginning of the closing-phase. First we will estimate the aspiration noise, and then we will obtain the noise envelope to be used as modulating function.

We will extract the noise component from the glottal parametrization error between (2) and (5) as:

$$r_{white} = h_{white} \star (g_{lf}(n) - g_{if}(n)), \quad (6)$$

where h_{white} is a whitening filter used to eliminate the distortion mainly occurring in the lower part of the spectrum due to modeling errors, and \star denotes convolution. We performed the whitening step using a 4th order filter derived by linear prediction analysis, followed by a high-pass filter with a cut-off frequency of 1kHz. The resulting residual waveforms resemble more closely those predicted by turbulent noise theory.

The next step is to extract the noise envelope, and to parametrize it using as few parameters as possible while retaining good resynthesis quality. We achieve this using the Hilbert transform, and then applying a low-pass filter to eliminate spurious components:

$$r_{env}(n) = \sqrt{\tilde{r}_{white}(n) \cdot \tilde{r}_{white}^*(n)}, \quad (7)$$

where $\tilde{\cdot}$ denotes the Hilbert transform, and $*$ denotes the conjugate. The noise envelope is parametrized as:

$$\hat{r}_{env}(n) = b_{lvl} + w_{lvl} \cdot W_{Han}^{sym}(n)|_{w_c, w_l} \quad (8)$$

where b_{lvl} accounts for the constant leakage present during the whole glottal cycle, W_{Han}^{sym} is a symmetric Hanning window, centered in w_c , width w_l and amplitude w_{lvl} . Figure 1 shows the aspiration noise and estimated noise envelope for a glottal period. We estimate the parameters by using least-squares to minimize the norm of the error between the extracted (7) and parametrized (8) envelopes:

$$\theta_{an}^{opt} = \arg \min_{\theta_{an}} \|\hat{\mathbf{r}}_{env}(\theta_{an}) - \tilde{\mathbf{r}}_{env}\|_2^2, \quad (9)$$

where $\theta_{an} = [b_{lvl}, w_{lvl}, w_c, w_l]^t$ is the vector of parameters. The synthetic aspiration noise is generated by modulating high-pass filtered AWGN with the fitted envelope \hat{r}_{env} , on a period-by-period basis.

3. Voice conversion

Generally, a voice conversion system may be divided in three components: a model of the acoustic space with a structure by classes, an acoustic classification machine and a mapping function. We will first explain the voice conversion system used in this work, and then we will detail the required modifications that are necessary in order to include our proposed parametrization.

3.1. Decision tree based voice conversion

Our VC algorithm [5] uses a classification and regression tree (CART) to classify the vocal tract data into phonetic categories. Then for each category, a Gaussian mixture model (GMM) is used to model the pdf of the training data, and to build the transformation function. Decision trees allow working with numerical data (such as spectral and glottal features) as well

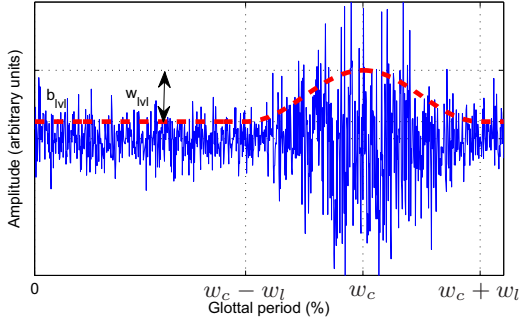


Figure 1: Proposed scheme to parametrize the glottal residual envelope using the model from eq. (8).

as categorical data (such as phonetic features) when building an acoustic model. The phonetic characteristics used are: the vowel/glide/consonant category, the point and manner of articulation for consonants, the height and the backness for vowels and glides, and the voicing.

The available training data is divided into two sets: the training set and the validation set. A joint GMM based conversion system [10] is estimated from the training set for the parent node t (the root node in the first iteration), and an error index $E(t)$ for all the elements of the training set belonging to that node is calculated:

$$E(t) = \frac{1}{|t|} \sum_{n=0}^{|t|-1} D(\tilde{\mathbf{y}}_n, \mathbf{y}_n), \quad (10)$$

where $|t|$ is the number of frames in the node t , \mathbf{y} is a target frame and $\tilde{\mathbf{y}}$ its corresponding converted frame. $D(\tilde{\mathbf{y}}, \mathbf{y})$ is a measure of the distance between target and converted frames.

All the possible questions of the set Q are evaluated at node t and two child nodes t_L and t_R are populated for each question q . The left descendant node t_L is formed by all the frames which fulfill the question and the right t_R node by the rest. The set Q is formed by binary questions of the form $is \{\tilde{\mathbf{y}} \in A\}$, where A represents a phonetic characteristic of the frame $\tilde{\mathbf{y}}$. For each child node, a joint GMM conversion system is estimated, and the error figures $E(t_L, q)$ and $E(t_R, q)$ for the training vectors corresponding to the child nodes t_L and t_R obtained from the question q are calculated. The increment of the accuracy for the question q at the node t can be calculated as:

$$\Delta(t, q) = E(t) - \frac{E(t_L, q)|t_L| + E(t_R, q)|t_R|}{|t_L| + |t_R|}. \quad (11)$$

The increment of accuracy for the training set is evaluated for each question and the question q^* corresponding to the maximum increment is selected. The node is split if the validation set accuracy for question q^* also increases. The tree is grown until there is no node candidate to be split. In order to avoid over-fitting to the training data, the tree is pruned using either a pre- or post-pruning method as explained in [5]. This is done independently for each source-target combination to find the optimal tree corresponding to each case.

To estimate a conversion function for each leaf, all the available data (training set plus validation set) is classified by the tree. Then, the data of each class is used to estimate a joint

GMM with Q component and the transformation function related is derived as [10]:

$$F(\mathbf{x}) = \sum_{q=1}^Q c_q(\mathbf{x}) \left(\mu_q^y + \sum_q^{y^x} \Sigma_q^{x^x-1} (\mathbf{x} - \mu_q^x) \right). \quad (12)$$

New source vectors are classified into leafs according to their phonetic features by the decision tree, and then converted according to the GMM based system belonging to its leaf. This is only applied to voiced segments, unvoiced segments are used unmodified.

3.2. Baseline parametrization

The baseline parametrization uses line spectral frequencies (LSF) to model the vocal-tract, derived using linear prediction (LPC) analysis. The distance $D(\tilde{\mathbf{y}}, \mathbf{y})$ used to compute the error index from eq. 10 is the mean of the Inverse Harmonic Mean Distance [11]:

$$D(\tilde{\mathbf{y}}, \mathbf{y}) = \sqrt{\sum_{p=1}^P c(p) (\tilde{y}(p) - y(p))^2} \quad (13)$$

$$c(p) = \frac{1}{w(p) - w(p-1)} + \frac{1}{w(p+1) - w(p)} \quad (14)$$

with $w(0) = 0$, $w(P+1) = \pi$ and $w(p) = \tilde{y}(p)$ or $w(p) = y(p)$ so that $c(p)$ is maximized (p is the vector dimension), weights more the mismatch in spectral picks than the mismatch in spectral valleys when working with LSF vectors.

In this case, we do not assume any model for the residual. To complete the conversion from the source speaker to the target speaker, a target LPC residual signal is predicted from the converted LSF envelopes as detailed in [5].

3.3. Proposed parametrization

We include our proposed parametrization by using the procedure explained above to grow three separate CART, one for each of the parameter sets (θ_{vt} , θ_{lf} and θ_{an} from section 2, i.e., vocal tract, glottal source and aspiration noise respectively). For the vocal tract CART, the IHMD (13) is used to compute the error (10). For the glottal source and aspiration noise CART conversion systems, we use the Euclidean distance between the converted and target vectors ($c(p) = 1$ in (13)).

4. Evaluation

As part of the TC-STAR project, UPC produced the language resources for supporting the evaluation of English/Spanish voice conversion. Four bilingual speakers English/Spanish recorded around 200 phonetically rich sentences in each language, using a mimic style to facilitate the alignment [10]. Ten sentences were reserved for testing, the rest were used for training. The recordings are of high quality (96kHz, 24 bits, three channels, including laryngograph), as explained in [12]. In this work, only the Spanish data set has been used, two female ($f1$ and $f2$) and two male ($m1$ and $m2$) voices, and four different source-target pairs have been trained ($m1$ to $m2$, $m1$ to $f2$, $f1$ to $m2$, and $f1$ to $f2$).

The evaluation was based on subjective rating by 14 human judges. As usual when evaluation VC algorithms, two metrics were used: one for rating the success of the transformation in achieving the desired speaker identification, and one for rating the quality. This is needed since strong changes usually achieve

the desired identity at the penalty of degrading the quality of the signal. The human judges were presented with examples from the transformed speech and the target one, and they had to decide on the similarity of the converted and target voices using a 5-point MOS scale (1 – completely different, to 5 – identical), and on the transformed voice quality using a 5-points MOS scale (1 – bad, 5 – excellent). Some natural source-target and target-target examples were also presented as a reference. The participants in this evaluation ignore the origin of the samples they are being presented with.

	Reference	Proposed	Orig Src	Orig Tgt
Quality	2.11	2.47	4.88	4.96
Similarity	2.87	3.02	1.62	4.79

Table 1: Evaluation results of the voice conversion system.

Table 1 presents the results of the similarity and quality tests, for both the baseline and proposed parametrizations. As we can see, there is a noticeable improvement in terms of transformed voice quality as a result of the new features used, which rises from 2.11 to 2.47 points. The speaker identity transformation is also rated as more successful (3.02 vs 2.87). From the last two columns, we can see that the original source and target speaker voices are judged to be different (rated 1.62), while the real target–target combinations are naturally judged identical. Real samples are found to have an excellent quality (4.87).

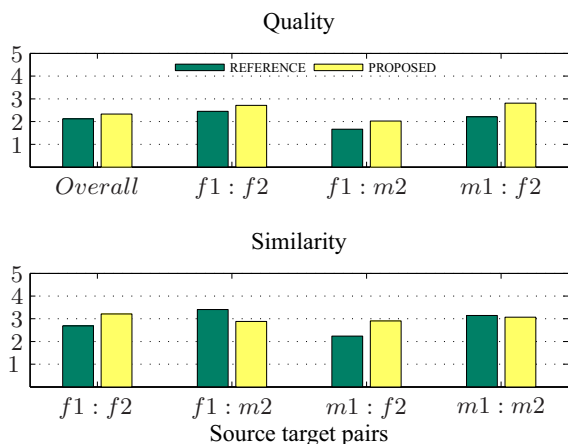


Figure 2: MOS evaluation results per conversion pair

Figure 2 contains the results separated per source-target pairs. As we can see, the proposed parametrization results in an improved quality in all four conversion directions. In terms of similarity, we observe that the transformation towards the second female voice $f2$ is more successful using the proposed parametrization, whereas in the other two cases there is a slight decrease in the performance.

5. Conclusions

Voice conversion (VC) technology transforms the voice of a source speaker so that it is perceived as that of a target speaker. This paper studies the inclusion of glottal source characteristics in voice conversion systems. We use our previously reported glottal analysis algorithm to obtain three sets of param-

eters: one for the vocal tract using LSF, another for the glottal source using the LF model, and a last one for the aspiration noise using a parametrized envelope to modulate in amplitude high-pass filtered AWGN noise. To evaluate the benefits of this new parametrization in voice conversion tasks, we use a reference conversion system that estimates a linear transformation function using a joint target/source model obtained with CART and GMM. The reference system is based on the LPC model, uses LSF to represent the vocal tract and a selection technique for the residual. To include the new parametrization, we use the reference system algorithm to build a VC system for each of the three parameter sets using CART and GMM.

We compared both parametrizations in the framework of an intra-lingual voice conversion task in Spanish. The tests show that the new source/filter representation clearly improves the overall performance, both in terms of speaker identity transformation and voice quality of the converted voice. However, the quality is still poor compared to that of equivalent speech synthesis systems. We are currently working in improving the modeling of the aspiration noise, which we believe could improve the overall synthetic quality. We will also investigate better conversion paradigms using our proposed parametrization.

6. References

- [1] G. Fant, *Acoustic theory of speech production*, 2nd ed. The Hague: Mouton, 1970.
- [2] P. Alku, “Parameterisation methods of the glottal flow estimated by inverse filtering,” in *VOQUAL*, Geneva, August 2003, pp. 81–87.
- [3] Q. Fu and P. Murphy, “Adaptive inverse filtering for high accuracy estimation of the glottal source,” in *ITRW on Non-Linear Speech Processing (NOLISP 03)*, Le Croisic, France, May 2003.
- [4] J. Pérez and A. Bonafonte, “Towards robust glottal source modeling,” in *INTERSPEECH*, Brighton, U.K., 2009, pp. 68–71.
- [5] H. Duxans, “Voice conversion applied to text-to-speech systems,” Ph.D. dissertation, Universitat Politècnica de Catalunya, 2006.
- [6] K. Yutani, Y. Uto, Y. Nankaku, A. Lee, and K. Tokuda, “Voice conversion based on simultaneous modelling of spectrum and f_0 ,” *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 0, pp. 3897–3900, 2009.
- [7] D. H. Klatt and L. C. Klatt, “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” *J. Acoust. Soc. Am.*, vol. 87, no. 2, February 1990.
- [8] G. Fant, A. Kruckenberg, J. Liljencrants, and M. Båvegård, “Voice source parameters in continuous speech. transformation of lf-parameters,” in *Proc. of the ICSLP*, Yokohama, 1994, pp. 1451–1454.
- [9] P. R. Cook, “Synthesis of the singing voice using a waveguide articulatory vocal tract model,” Ph.D. dissertation, Stanford University, 1991.
- [10] A. Kain, “High resolution voice transformation,” Ph.D. dissertation, OGI School of Science and Engineering, 2001.
- [11] R. Laroia, N. Phamdo, and N. Farvardin, “Robust and efficient quantization of speech LSP parameters using structured vector quantizers,” in *International Conference on Acoustics, Speech, and Signal Processing*, 1991, pp. 641–644.
- [12] A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. van den Heuvel, H. Hain, X. S. Wang, and M. N. Garcia, “TC-STAR: Specifications of language resources and evaluation for speech synthesis,” in *LREC*, Genoa, Italy, 2006.