

The Linear Transformation of LF Glottal Waveforms for Voice Conversion

Arantza del Pozo and Steve Young

Cambridge University Engineering Department
Trumpington Street, Cambridge, England, CB2 1PZ
ad371@eng.cam.ac.uk, sjy@eng.cam.ac.uk

Abstract

Most Voice Conversion (VC) systems exploit source-filter decomposition based on linear prediction (LP) to transform spectral envelopes, incurring as a result various issues related to the oversimplification of the LP voice source model. Whilst residual prediction methods can mitigate this problem, they cannot be used to modify voice source quality. In this paper, a system which employs linear transformations to convert both the spectral envelope and the LF glottal waveform is presented. Its performance is shown to be comparable to that of a state-of-the-art VC implementation in terms of speaker identity conversion but its output has better quality. In addition, it is also capable of transforming the quality of the voice source.

Index Terms: LF waveform, deconvolution, voice conversion

1. Introduction

The most widely used speech signal representations are the Source-Filter Model and the Sinusoidal Model. The Source-Filter representation [1] is based on a simple production model composed of a glottal source waveform exciting a time-varying filter loaded at its output by the radiation of the lips. The main challenge in Source-Filter modelling is the estimation of the glottal waveform and vocal tract filter parameters from the speech signal.

Linear Prediction (LP) is one popular technique used to obtain a combined parameterisation of the glottal source, vocal tract and lip radiation components in a unique all-pole filter $H(z)$. Such a filter is then excited, as shown in Figure 1, by a sequence of impulses spaced at the fundamental period T_0 during voiced speech and by white gaussian noise during unvoiced speech. If the speech signal were truly the response of an all-pole filter, the LP error or residual would be a train of impulses spaced at the voiced excitation instants and the impulse/noise voice source modelling would be accurate. In practice, however, the LP residual looks more like a white noise signal with larger values around the instants of excitation. While exciting the LP filter with the LP residual results in speech that is indistinguishable from the original, using an impulse train as the voiced excitation produces speech with a very buzzy quality. The strength of LP lies in its ability to automatically estimate a set of filter coefficients which compactly represent the envelope of the speech spectrum, making it popular in applications where the spectral characteristics of the speech wave need to be captured with a small number of parameters. Its main drawback, on the other hand, stems from the over-simplified modelling of the glottal source which prevents its use in systems requiring high-quality speech outputs.

Sinusoidal Models assume the speech waveform to be composed of the sum of a small number of sinusoids with time-varying amplitudes, frequencies and phases. Such modelling

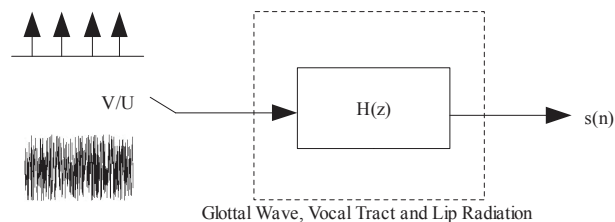


Figure 1: Schematic diagram of the LP model

was mainly developed by McAulay and Quatieri [2] in the mid-1980's and has been shown to be capable of producing high-quality speech even after pitch and time-scale transformations. However, because of the high number of sinusoidal amplitudes, frequencies and phases involved, sinusoidal modelling results less flexible than the source-filter representation to modify spectral features.

In order to obtain high-quality converted speech, state-of-the-art VC implementations mainly employ variations and extensions of the original sinusoidal model. In addition, they generally adopt a source-filter formulation based on LP to carry out spectral transformations. Unfortunately, this implies the use of LP residuals as the voice source representation. Sinusoidal VC systems have developed residual prediction and selection methods [3] based on the correlation between spectral envelope and LP residuals to reintroduce the target spectral detail lost after envelope conversion. Because residuals contain the errors introduced by the LP parameterisation, residual prediction techniques have been found to improve conversion performance. However, LP residuals do not constitute an accurate model of the voice source and residual prediction alone is not capable of modifying the quality of the voice source. This prevents their use in applications requiring voice quality modifications such as, for example, speech repair.

In this paper, a source-filter modelling formulation which uses a representation of the glottal source more accurate than LP residuals is adopted for voice conversion. This allows the use of linear transformations for the conversion of the voice source.

2. Joint Estimation Analysis Synthesis

The Joint Estimation Analysis Synthesis (JEAS) model used for the analysis, modification and synthesis of speech is illustrated in Figure 2. It follows the general Source-Filter representation introduced in Section 1, employing white gaussian and amplitude-modulated white gaussian noise to model the turbulence and aspiration noise components respectively, a digital differentiator for lip radiation and an all-pole filter to represent the vocal tract. However, instead of simplifying the mod-

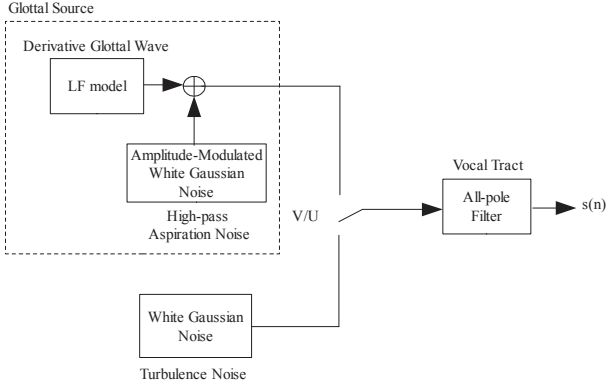


Figure 2: Schematic diagram of the JEAS Model

elling of the voice source with a two-pole filter as in LP, the Liljencrants-Fant (LF) model [4] is adopted to better capture the characteristics of the derivative glottal wave. Then, in order to estimate the different model component parameterisations from the speech wave, it applies the joint voice source and vocal tract parameter estimation technique based on Convex Optimization proposed in [5].

2.1. Modelling the Voice Source: LF model

Among the existing glottal wave parameterisations, the LF model [4] has become the model of choice for research on the glottal source. It has been shown to be capable of modelling a wide range of naturally occurring phonations and the effects of its parameter variations are well understood. It exploits the linearity and time-invariance properties of the Source-Filter representation and assumes the commutation of the vocal tract and lip radiation filters to combine the modelling of the source excitation and lip radiation in the parameterisation of the derivative of the glottal waveform. Typical LF pulses corresponding to glottal and derivative glottal waves are shown in Figure 3. Mathematically, it can be described as:

$$g(n) = \begin{cases} E_0 e^{\alpha n} \sin(\omega_g n) & 0 \leq n < T_e \\ -\frac{E_e}{\epsilon T_a} [e^{-\epsilon(n-T_e)} - e^{\epsilon(T_c-T_e)}] & T_e \leq n < T_c \end{cases} \quad (1)$$

Along with E_e , the LF pulse can be uniquely determined by the timing parameters: (T_p, T_e, T_a, T_c) . These parameters can be easily identified from the estimated derivative glottal wave. Therefore, they are generally obtained first and the synthesis parameters $(E_0, \alpha, \omega_g, \epsilon)$, from which the LF waveform can be computed directly are then derived. Another important set of LF parameters are the R-parameters (R_g, R_k, R_a) , which are normalised respect to T_0 and correlate with the most salient glottal phenomena, i.e. the glottal pulse width and the skewness and abruptness of closure.

$$R_g = \frac{T_0}{2 \cdot T_p}; R_k = \frac{T_e - T_p}{T_p}; R_a = \frac{T_a}{T_0} \quad (2)$$

2.2. Joint Source-Filter Deconvolution

The method employed to obtain the JEAS voice source and vocal tract model parameters from the speech wave is based on the joint estimation approach proposed in [5]. It involves using a voice source model simple enough to allow the source-filter deconvolution to be formulated as a Convex Optimization

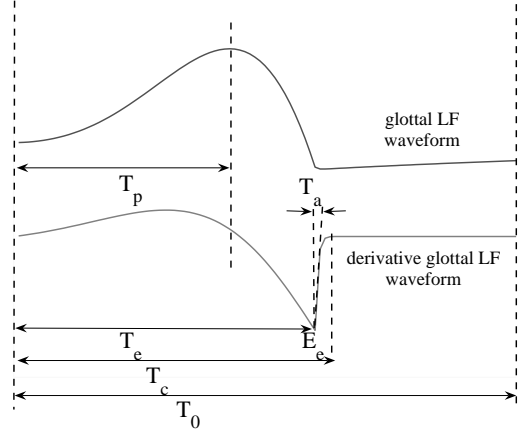


Figure 3: Typical glottal and derivative glottal LF waveforms and identifying features: glottal excitation strength E_e and timing parameters (T_p, T_e, T_a, T_c)

problem. Then, the derivative glottal waveform obtained by inverse filtering (IF) with the estimated filter coefficients is reparameterised by LF model fitting.

Deconvolution is accomplished by minimising the squared error between the modelled Rosenberg-Klatt (RK) and the true derivative glottal waveforms. The RK model consists of the basic voicing waveform of Equation (3) and a low-pass filter, $\frac{1}{1-\mu z^{-1}}$, which encodes the spectral tilt.

$$\hat{g}(n) = \begin{cases} 0 & 1 \leq n < n_c \\ 2a(n - n_c) - 3b(n - n_c)^2 & n_c \leq n < T_0 \end{cases} \quad (3)$$

where n_c represents the duration of the closed phase and the parameters a and b hold $a = b \cdot (T_0 - n_c) \cdot T_0$.

The true derivative glottal wave $g(n)$ can be defined as

$$g(n) = s(n) - \sum_{k=1}^p \alpha_k s(n - k) \quad (4)$$

where $s(n)$ is the speech wave and α_k are the coefficients of the vocal tract all-pole filter.

The error between the modelled and the true derivative glottal waves $e(n)$ can be calculated by subtracting Equations (3) and (4) in the closed and open phases

$$e(n) = \hat{g}(n) - g(n) = \begin{cases} 0 - s(n) + \sum_{k=1}^p \alpha_k s(n - k) \\ 2a(n - n_c) - 3b(n - n_c)^2 - s(n) + \sum_{k=1}^p \alpha_k s(n - k) \end{cases} \quad (5)$$

which rearranged and rewritten in matrix form gives

$$E = FX - S \quad (6)$$

where $X = [\alpha_1 \ \dots \ \alpha_p \ a \ b]'$ is the parameter vector to estimate. [5] demonstrated the least squares error optimization of Equation (6) to be convex and thus, efficiently solvable via Quadratic Programming. In order to smooth possible adjacent parameter discontinuities, linear interpolation of source and tract parameter trajectories is applied.

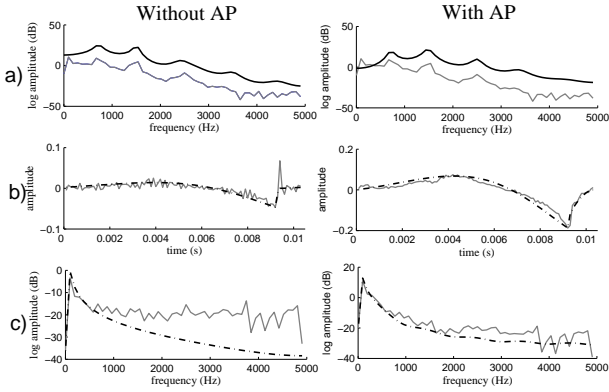


Figure 4: Effect of adaptive pre-emphasis: a) speech spectrum and estimated spectral envelope; b) IF derivative glottal wave and fitted LF waveform; c) IF derivative glottal wave spectrum and fitted LF wave spectrum

Whilst in [5] the low-pass filter representing the spectral tilt is separated from the source and incorporated into the vocal tract to allow the formulation of the convex optimization problem, the spectral tilt is encoded differently in our implementation. We use adaptive pre-emphasis (AP) to estimate and remove the spectral tilt filter contribution from the speech wave before convex optimization. The effect of adaptive pre-emphasis is illustrated in Figure 4. The vocal tract filter envelope estimates obtained this way do not encode source spectral tilt characteristics, which are reflected in the closing phase of the resulting derivative glottal waveforms instead. This has been found to improve the fitting of the return phase of the LF model and thus, of the high frequencies of the glottal source.

As in [5], wavelet denoising is used to extract the glottal aspiration noise component from the IF derivative glottal wave estimate. However, it is modelled differently: by modulating zero mean unit variance Gaussian noise with the LF waveform fitted for each pitch period and adjusting its energy to match that of the aspiration noise estimate N_e .

Visual inspection of the estimated spectral envelope and fitted LF waveforms has shown that the joint source-filter deconvolution technique is successful and does not introduce noticeable artifacts in the parameterisation. Moreover, speech resynthesised with the estimated JEAS model parameters has been found to be almost indistinguishable from the original.

3. Linear Transformations for LF Glottal Waveform Conversion

Linear transformations provide a robust and efficient method of spectral envelope conversion. The parameterisation of the glottal source provided by JEAS Modelling allows the same robust transformation technique to be applied to the voice source parameters, avoiding the need to predict appropriate LP residuals.

Five-dimensional feature vectors derived from the JEAS model parameters linked to the voice source of every pitch period have been employed for glottal waveform conversion. Each feature vector is composed of the glottal excitation strength E_e , the normalised R-parameters (R_g, R_k, R_a) and the energy of the aspiration noise N_e .

As proposed for spectral envelope conversion in [6], Gaussian Mixture Models (GMMs) can also be used to describe the source and target glottal feature spaces, classify them into

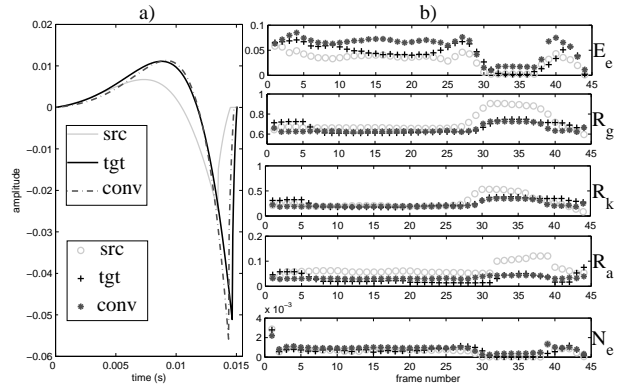


Figure 5: Linear Transformation of LF Glottal Waveforms: a) source, target and converted derivative glottal LF waves; b) source, target and converted trajectories of the glottal feature vector parameters (E_e, R_g, R_k, R_a, N_e)

M classes and train class specific linear transformations. A weighted sum of the linear transformations can then be employed to convert each glottal source feature vector x

$$\mathcal{F}(x) = \left(\sum_{m=1}^M \lambda_m(x) W_m \right) \bar{x} \quad (7)$$

where \bar{x} is the extended feature vector $\bar{x} = [x', 1]'$ and λ_m is the interpolation weight of transformation matrix W_m , its value given by the probability of vector x belonging to class C_m

$$\lambda_m(x) = P(C_m|x) = \frac{\alpha_m N(y_i; \mu_m, \Sigma_m)}{\sum_{i=1}^M \alpha_i N(y_i; \mu_i, \Sigma_i)} \quad (8)$$

α_m, μ_m and Σ_m being the weights, means and variances of the GMM components respectively and $N()$ representing the Normal Distribution.

The transformation matrices W_m can be estimated using parallel training data and a least square error criterion [6].

As it can be seen in Figure 5, glottal conversion does move the source feature vector parameter contours closer to the target and produces converted glottal waveforms which are more similar to the target.

4. Evaluation

In order to evaluate the glottal waveform transformation technique, its performance was compared to the high-quality voice morphing system described in [6]. This employs a Pitch-Synchronous Harmonic Model (PSHM) to represent and manipulate the speech signal, linear transformations to convert spectral envelopes, a prediction method to transform residuals and a phase prediction technique to mitigate the artifacts caused by the unnatural sinusoidal phase dispersion. In contrast, the JEAS VC system uses linear transformations to convert both the spectral envelopes and the LF glottal waveforms.

A conversion task based on the VOICES database [7] involving male-to-male (MM), male-to-female (MF), female-to-male (FM) and female-to-female (FF) transformations was used for the evaluation. Of the 150 parallel sentences available per conversion experiment, the first 120 were used for training and the remaining 30 for testing. Spectral vectors of order 30 were employed to train 8 linear spectral envelope transforms

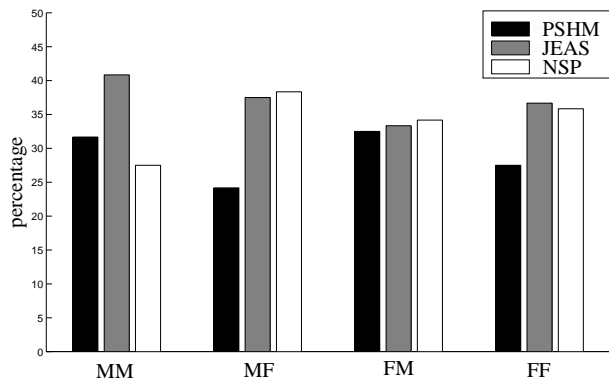


Figure 6: Results of the ABX test

between each source and target speaker pair. Aligned source-target vector pairs were obtained by applying forced alignment to mark sub-phone boundaries and using Dynamic Time Warping (DTW) to further constrain their alignment. For PSHM residual and phase prediction, target GMMs and codebooks of 40 classes and entries were built. For JEAS, glottal waveform conversions were also applied using 8 linear transforms per pair. These numbers were the optimal for the experimental setting.

A listening test was carried out to assess the performance of the PSHM and JEAS systems in terms of recognizability and quality. 12 subjects took part in the perceptual study.

In the first part an ABX test in which subjects were presented with PSHM-converted (A), JEAS-converted (B) and target (X) utterances and were asked to choose the speech sample A or B they found sounded more like the target X in terms of speaker identity. The prosody of the target was employed to synthesise the converted sentences in order to normalise the pitch, duration and energy differences between speakers for the perceptual comparison. 10 utterances of each conversion type (MM, MF, FM, FF) were presented. The order of the samples in terms of conversion type and conversion system was randomised. Informal listening of the utterances transformed using the PSHM and JEAS conversion systems revealed that it was often very difficult to convincingly choose between systems in terms of speaker identity. For this reason, subjects were also allowed to select a 'NO STRONG PREFERENCE' option when they did not have a strong preference towards one of the presented A or B speech samples.

Figure 6 shows the results of the ABX test. The JEAS-converted samples were preferred over the PSHM-converted ones overall, which suggests that glottal source transformation slightly improves speaker identity conversion. However, the 'NO STRONG PREFERENCE' (NSP) option was selected almost as often as the JEAS-converted utterances in general, which reveals that subjects often found it difficult to distinguish between conversion systems in terms of speaker identity. Presumably this is because the most important speaker identifying cues, i.e. the spectral envelopes, were transformed using the same method in the two cases. Hence, it is expected that both systems should perform equally in terms of speaker recognizability. Overall, the results show that the residual prediction and glottal waveform conversion techniques are comparable in terms of perceptual speaker identity transformation.

The second part aimed at determining which system produces speech with a higher quality. Subjects were presented with PSHM and JEAS converted speech utterance pairs and

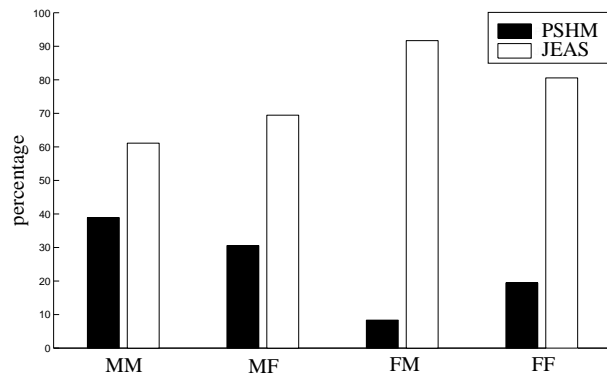


Figure 7: Results of the quality comparison test

asked to choose the one they thought had a better speech quality. The results are shown in Figure 7. There is a clear preference for the sentences converted using the JEAS method which was chosen 75.7% of the time on average. There was thus a clearly distinguishable quality difference between the PSHM and JEAS transformed samples. Utterances obtained after PSHM conversion have a 'noisy' quality caused by phase discontinuities. Whereas, the JEAS converted sentences sounded much smoother. This quality difference might also have favoured the slight preference towards JEAS conversion in the ABX test.

5. Conclusions

In this paper a new method to convert glottal source characteristics in VC applications has been presented. It involves the application of linear transformations to convert LF waveforms obtained using a speech representation capable of automatically estimating glottal source and vocal tract parameterisations from the speech wave. The proposed VC implementation achieves conversions comparable to the state-of-the-art in terms of speaker recognizability but with a higher speech quality.

6. Acknowledgements

This work was supported by a researcher training grant from the Government of the Basque Country. The authors thank the volunteers of the perceptual tests for their assistance.

7. References

- [1] G. Fant, *Acoustic Theory of Speech Production*. The Netherlands: Mouton-The Hague, 1960.
- [2] R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 744–754, 1986.
- [3] D. Suendermann, A. Bonafonte, H. Ney, and H. Hoega, "A study on residual prediction techniques for voice conversion," in *Proc. ICASSP*, 2005, pp. 13–16.
- [4] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, Tech. Rep., 1985.
- [5] H. Lu, "Toward a High-Quality Singing Synthesizer with Vocal Texture Control," Ph.D. dissertation, Stanford University, 2002.
- [6] H. Ye and S. Young, "Quality-enhanced Voice Morphing using Maximum Likelihood Transformations," *IEEE Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1301–1312, 2006.
- [7] A. Kain, "High Resolution voice transformation," Ph.D. dissertation, Oregon Health and Science University, 2001.