Exploiting time and frequency domain measures for precise voice source parameterisation

John Kane, Irena Yanushevskaya, Ailbhe Ní Chasaide, Christer Gobl

Phonetics and Speech Laboratory,
Centre for Language and Communication Studies
Trinity College Dublin
kanejo@tcd.ie, yanushei@tcd.ie, anichsid@tcd.ie, cegobl@tcd.ie

Abstract

Much of our research has focused on the role of the voice source in the prosody of spoken language, including its linguistic and expressive dimensions. However, as automatic methods, both for deriving the voice source and for modelling it tend to lack robustness, we have generally conducted studies on small amounts of speech data. These studies have involved the use of labour intensive methods which require pulse-by-pulse manual fine-tuning. This paper describes a method to model the voice source automatically by taking into account some of the strategies involved in the manual fine-tuning approach. The method combines exhaustive search, dynamic programming and optimisation methods to overcome the known difficulties of standard automatic algorithms. A quantitative evaluation revealed parameter values for the proposed method that were closer to the reference values, than those obtained using a standard time-based method.

Index Terms: voice source, LF model, parameterisation, prosody, dynamic programming.

1. Introduction

A major research strand at the Phonetics and Speech Laboratory in Trinity College Dublin concerns the role of the voice source in speech. This research includes both descriptive studies and the development of more robust analytic tools and methodologies. One goal is to describe the prosody of the voice, i.e. how dynamic, temporal variation of the entire voice source (f_0 and phonation quality), provides both the underlying linguistic prosody as well as its expressive dimension. As part of this endeavour we have been looking at the source correlates of focus and deaccentuation [1]. These studies have involved the use of labour intensive methods which require pulse-by-pulse manual fine-tuning. This paper looks at methodological developments, drawing on data where utterances were elicited with different focal accentuation patterns.We present a method for voice source parameterisation that allows us to overcome some of the difficulties that arise with standard automatic analysis methods.

In order to derive an estimate of the voice source we first consider the speech production process (in the frequency domain) as:

$$S(f) = G(f)V(f)L(f) \quad (1)$$

where the spectrum of the speech output, $S(f)$, is the product of the three factors $G(f)$, $V(f)$ and $L(f)$, where $G(f)$ is the spectrum of the glottal flow signal (i.e. the voice source spectrum), $V(f)$ is the transfer function of the vocal tract, $L(f)$ is the spectral effect of sound radiation at the lip opening and f is frequency in Hz. In the time domain, the effect of radiation at the lips is typically modelled as a first order differentiator, in which case Eq. (1) can be reduced to:

$$S(f) = G_{diff}(f)V(f) \quad (2)$$

where G_{diff} is the spectrum of the differentiated glottal flow. Thus, the voice source signal can be obtained through inverse filtering if the vocal tract transfer function is known. However, $V(f)$ is not directly observable and as a result this inverse filtering operation becomes an immensely difficult signal processing task. Many automatic algorithms exist for vocal tract inverse filtering, including: closed-phase methods [2], iterative and adaptive linear predictive coding based methods [3], and methods which consider the mixed-phase properties of speech [4]. However, inverse filtering is still generally considered to be an unsolved problem.

Due to the frequent problems of automatic algorithms, we tend to rely on an inverse filtering method which derives initial estimates using an automatic closed-phase inverse filtering technique followed by an optimisation procedure involving manual fine-tuning. The user modifies the estimated formant frequencies and bandwidths and utilises both time and frequency domain displays to obtain maximum formant cancellation [5]. The inverse filtering of the speech signal provides an estimate of the voice source, which we then characterise by fitting the LF source model to each individual glottal pulse, thus facilitating the parameterisation of important features in the source signal. Again, an automatic algorithm is first used to derive an initial model fit, which is followed by manual fine-tuning to get an improved fit. As with the inverse filtering, the user is visually guided to ensure optimisation in both the time and frequency domain. Furthermore, the user ensures that subsequent model pulses do not have unwarranted discontinuities. The manual fine-tuning involved in both the inverse filtering and the source parameterisation is extremely labour intensive. However, due to the limitations of current automatic algorithms, we have found this approach necessary if a precise description of the voice source is required.

In this paper we describe an automatic voice source parameterisation method which attempts to simulate some of the strategies used by the researcher when applying the visually guided optimisation of the model fit. To do this, we initially use an exhaustive search method which provides the N best parameter settings for the model fit, in terms of both time and frequency domain criteria. A dynamic programming algorithm is then used to select the optimal path of parameter values by considering both the 'target cost' (i.e. the temporal and spectral fit of the modelled pulses) and the 'transition cost' (i.e. the

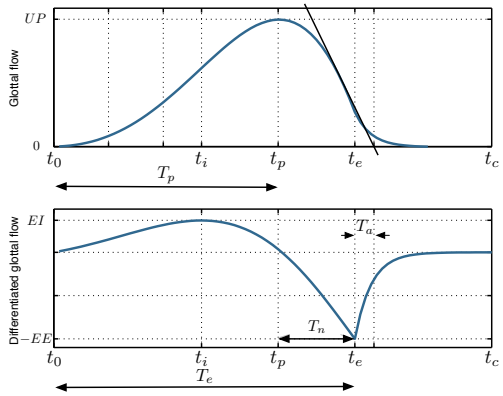

Figure 1: *Example LF model glottal flow (top) and differentiated glottal flow (bottom) pulse.*

continuity in the parameter trajectories of the modelled glottal pulses). To evaluate the new method, we compared its performance to that of a standard automatic algorithm based on model fitting in the time domain.

2. Proposed method

We here describe a method for estimating LF model parameter values by considering some of the information typically available to a researcher using a manual fine-tuning approach, i.e. time domain and frequency domain information, and overall parameter trajectory. It is assumed that the voice source signal has been derived beforehand using either an automatic method or using a manual fine-tuning method (as is used here).

2.1. LF model

Our studies of the voice source have typically involved the use of the Liljencrants-Fant (LF) model (see Fig. 1) to characterise the salient voice source characteristics [6]. The LF model is a 5 parameter model (including f_0 and assuming $t_c = t_o$ of the following pulse) of differentiated glottal flow (i.e. the voice source). The shape of the LF model can be characterised by the three R-parameters, calculated as follows:

$$Rg = \frac{T_0}{2T_p} \quad (3)$$

$$Rk = \frac{t_e - t_p}{T_p} \quad (4)$$

$$Ra = \frac{T_a}{T_0} \quad (5)$$

where T_p is the duration from the time point of glottal opening to the time point of peak glottal flow amplitude (Fig. 1), t_e is the time point of the main excitation and T_a represents the effective duration of the return phase. A further R-parameter, Rd, was developed to provide a single parameter which captures most of the covariation of the LF model parameters [7], and is derived using:

$$Rd = 1000 \cdot \left(\frac{UP}{EE}\right) \cdot \left(\frac{f_0}{110}\right) \quad (6)$$

where UP is the peak amplitude of glottal flow and EE is the negative amplitude of the main excitation (differentiated glottal flow). The other R-parameters (Rg, Rk, Ra) can be predicted from Rd, following the regression analysis described in [7].

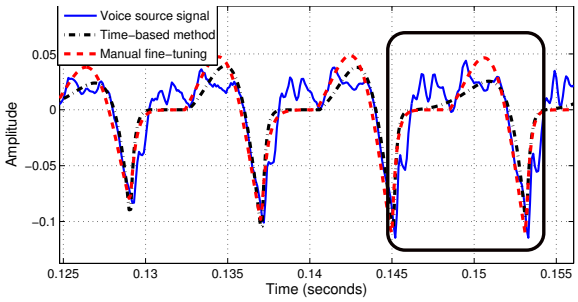

Figure 2: *Estimated voice source signal (solid line), a synthesised source signal using manually tuned parameters (dashed line) and a source signal derived using parameters estimated using the standard automatic time based parameterisation method (dot-dashed line). The change in the model setting in the final pulse highlights the potential for inconsistency.*

2.2. GCIs, f_0 and EE

In order to estimate f_0 and EE values from the voice source signal, we start by estimating the glottal closure instants (GCIs) using an adapted version of the method described in [8]. A mean-based signal is derived from the speech signal, $s(n)$, using:

$$y(n) = \frac{1}{2N+1} \sum_{m=-N}^{N} w(m)s(n+m) \quad (7)$$

where $w(m)$ is a Blackman window [4]. Peaks are then measured in the signal $y(n)$ and the regions between adjacent positive peaks are used as search regions. The location of the maximum negative amplitude in the voice source signal (differentiated glottal flow) within each search interval is chosen as the GCI location. The maximum negative amplitude in each search region is used as EE and f_0 is determined by the reciprocal of the duration between adjacent GCIs.

2.3. Exhaustive search over Rd

Standard automatic time domain approaches to LF model fitting typically involve estimating initial parameter values by direct measurements of the voice source pulse and then refining these estimates using an optimisation procedure. One common problem with this approach is that direct measurements can often yield poor initial parameter values. This is frequently due to inconsistency in marking the point of glottal opening, t_o. This is highlighted in Fig. 2 where the final pulse from the time-based method (dot-dashed line) changes shape considerably from the previous pulses. Subsequent use of an optimisation algorithm does little to rectify the problem. To overcome this we propose the use of an exhaustive search method which involves the generation and analysis of a wide range of LF-model parameter configurations, and saving those configurations that minimise a specific error function However, to cover the full range of possible LF model configurations would be computationally prohibitive. Therefore, rather than searching all R-parameter combinations, we simplify the search by varying only Rd. Rd is changed in steps of 0.1 within the range [0.3, 5].

The search is done by first taking a GCI centred frame of the voice source signal, $U_g'(n)$, and windowing it using a Hanning window. We use a frame length, L, of three local glottal periods to ensure clear harmonic peaks in the spectrum. We then measure the amplitude spectrum in dB from the windowed voice source segment. Harmonic amplitudes are measured in the spectrum up to a specific maximum frequency (H_{max}). In

this work we set H_{max} to 3 kHz. Then for each step in the search an LF model pulse is generated using f_0 and EE (previously calculated) and using Ra, Rk and Rg as derived from the current Rd value. A synthetic signal is obtained by concatenating the LF pulses together and getting a three pulse length segment again centred on a GCI. The spectrum and harmonics are measured as above. For each Rd, an error value is measured between the two harmonic sets using:

$$\text{spec\_err} = \left\{0.5 - |cor\{h_U(m), h_{LF(m)}\}|\right\} \cdot w_s \quad 1 \leq m \leq N \tag{8}$$

where h_U and h_{LF} are the harmonic amplitudes measured from the voice source signal and the synthesised LF model signal, respectively, N is the number of harmonics of frequencies below H_{max}, $cor\{.\}$ is the Pearson correlation between the harmonics h_U and h_{LF} and w_s is a constant weight (see Section 3.2). A time domain error value is measured using:

$$\text{time\_err} = \left\{0.5 - |cor\{U'_{gLF}, U'g(t)\}|\right\} \cdot w_t \tag{9}$$

where U'_{gLF} is a synthesised LF model source signal of length L and set using the current Rd value, t is the sample range from the start and end point of the current frame, and w_t is a constant weight. We then consider the N_{cand} (empirically set to 5) Rd values that minimise the total error function:

$$\text{total\_cost} = \text{spec\_err} + \text{time\_err} \tag{10}$$

2.4. Dynamic programming

We use a dynamic programming method to select the optimal path of Rd values through the input speech signal. The particular dynamic programming method used here is described in [9] and has been used in the popular *get_f0* pitch tracker.

We define the target cost, $d(i, j)$, as the error value calculated in the exhaustive search (Eq. 10) for each Rd candidate in each analysis frame, where $1 \leq j \leq N_{cand}$, $1 \leq i \leq M$ and M is the number of GCIs (i.e. the number of analysis frames). The transition cost is:

$$\delta_{i,j,k} = \{0.5 - cor\{seg_{i,j}, seg_{i-1,k}\}\} \cdot w_{tr} \tag{11}$$

where $seg_{i,j}$ refers to a single generated LF model pulse using the R-parameters predicted from the j-th Rd candidate at frame i and $seg_{i-1,k}$ refers to an LF pulse generated using the previously chosen Rd value. This transition cost is based on the observation that, like the vocal tract, voice source pulses should be reasonably slowly varying over a short timespan (e.g., 20 ms). This may not be the case for certain voice qualities (e.g., harsh and creaky voice). These voice qualities are not contained in the speech data used in the present work. An objective function is, hence, defined for a given frame i:

$$D_{i,j} = d_{i,j} + \min_{k \in N}\{D_{i-1,k} + \delta_{i,j,k}\}, \quad 1 \leq j \leq N_{cand} \tag{12}$$

which is initialised with: $D_{o,j} = 0, \quad 1 \leq j \leq N_{cand}$. The vector $q(i)$ is used to save the index of the optimal Rd (obtained by $\underset{j}{\text{argmin}}(D_{i,j})$ for $1 \leq i \leq M$).

2.5. Optimisation

Although the Rd parameter can be used to characterise much of the glottal pulse types arising in breathy to tense voice qualities, it is likely that some glottal pulses will exist outside the constraints of Rd. Furthermore, it is not our intention to reduce the degrees of freedom of the model. To overcome this we refine parameter values using an optimisation method. For each analysis frame we derive Ra, Rk and Rg from the Rd value, selected from the dynamic programming method. We then use a simplex-based method [10], which allows multi-variable optimisation. The 3 R-parameters are allowed to vary to minimise the same error function shown in Eq. (10).

3. Evaluation

3.1. Speech data

The speech data from 6 male speakers were used in our evaluation. Each speaker uttered the sentence *WE WERE aWAY a YEAR ago*, with narrow focus on each of the potentially accented syllables (*WE*, *WERE*, *-WAY* and *YEAR*) with both rising and falling pitch patterns. A broad focus and a deaccented rendition of the utterance were also recorded. Overall there were 10 sentences per speaker with the exception of one, from whom the 4 rising pitch utterances were not elicited. The speech samples for this speaker (6 utterances) were used for weight setting (see Section 3.2) and were subsequently excluded from the testing, leaving 50 utterances for evaluation.

Audio was captured in a semi-anechoic recording studio using high quality recording equipment (a B & K 4191 free-field microphone and a B & K 7749 pre-amplifier) and was digitised at 44.1 kHz (using a LYNX-two sound card), which was then downsampled to 10 kHz. The DC-component was removed using an 8th order high pass Butterworth filter with a cut-on frequency of 60 Hz. Filtering was carried out forwards and backwards to maintain the original phase spectrum of the signal.

All speech signals were inverse filtered by the second author using the manual fine-tuning software [5], where the user adjusts the initial formant frequencies and bandwidths for each analysis frame and uses time and frequency domain displays to achieve full formant cancellation.

3.2. Weight setting

In our evaluation, the proposed method is configured as described in Section 2. However, the weights w_s, w_t and w_{tr} need to be set. The setting of these weights is crucial for modelling the relative importance of different types of information used in the manual fine-tuning approach and, hence, they need to be set carefully.

Using the speech data from one speaker (see Section 3.1) an exhaustive search was conducted to test all combinations of the three weights in the range [0, 1] with a step of 0.1 (1331 possible combinations). Note that all three cost elements were designed to lie within the range [-0.5, 0.5]. For each combination, analysis was carried out on the 6 sentences and a synthetic voice source signal was generated using the extracted parameter values. This was compared to the voice source signal, generated using the reference parameter values, by calculating the Pearson correlation coefficient. The combination with the highest correlation score (averaged across the 6 sentences) was kept as the setting for the weights. The analysis resulted in the weights 0.6, 1 and 0.3 for w_s, w_t and w_{tr} respectively. This suggests that the manually fine-tuning user favours the time domain information for fitting the model. However frequency domain and transitioning information also carry importance.

3.3. Comparison algorithm

We use a standard time based LF model fitting algorithm as a baseline method in our evaluations. The algorithm is described in [11], and involves first estimating LF model parameters by

direct estimates from each voice source pulse. A multi-variable optimisation method [10] is then used to refine the fit by minimising a sum of squares error function.

3.4. Reference values and evaluation metrics

Objective evaluation of voice source parameterisation is a difficult task. Some researchers tend to use synthetic stimuli to provide a quantitative evaluation with known reference parameter values. However these signals may lack the very details that cause trouble for voice source parameterisation (e.g., the presence of aspiration noise). Others use EGG signals for obtaining reference values. It is not generally feasible, however, to obtain a full set of voice source parameter values from the EGG signal. In the present study we used parameter values obtained using a fine-tuning method [5] as our reference values. With an estimated voice source signal, initial LF model settings were derived for each glottal pulse. The second author, who is highly experienced with this type of analysis, then used our manual fine-tuning technique in order to ensure optimal fitting of the LF model to each of glottal pulses of the voice source signal.

Using these reference values, the performance of the proposed method was evaluated by comparing it to the performance of the baseline method (described in Section 3.3). We considered the three R-parameters, Rg, Rk and Ra, and used the relative error as an evaluation metric:

$$\text{Relative error} = \frac{|param_{ref} - param_{est}|}{param_{ref}} \qquad (13)$$

where $param_{ref}$ are the reference parameter values and $param_{est}$ are the parameter values as estimated by the automatic algorithms. We also calculated the Pearson correlation score for each utterance between the generated source signal using the reference values and the one generated using the estimated parameter values. This metric would be an indicator of the overall similarity between the automatic model fitting and that of our reference. Independent Student t-tests were used, separately for each parameter, to compare distributions of relative error scores across the two methods.

4. Results

The distributions of relative error scores for the three parameters are presented in Fig. 3. The proposed method produced significantly lower relative error scores for Rg [t = -3.7192, p < 0.001], Rk [t = -2.6885, p < 0.01] and Ra [t = -7.9259, p < 0.01]. Also, for Rg and Rk there is clearly lower variance in the error scores for the proposed method. This is likely due to the known difficulty of consistently marking the point of glottal opening, t_o, in the standard method (as shown in Fig. 2), which would consequently affect Rg and Rk values. Ra, on the other hand, is not affected by the position of t_o and as a result both methods display comparably low variance.

For the correlation scores comparing test signals with reference synthetic source signals, the proposed method produced a higher mean correlation score (R = 0.890) compared with the standard method (R = 0.795), and this difference was found to be significant [t = 3.2576, p < 0.01]. This implies that the voice source modelling of the proposed method is consistently more similar to the manual modelling method, than the standard method.

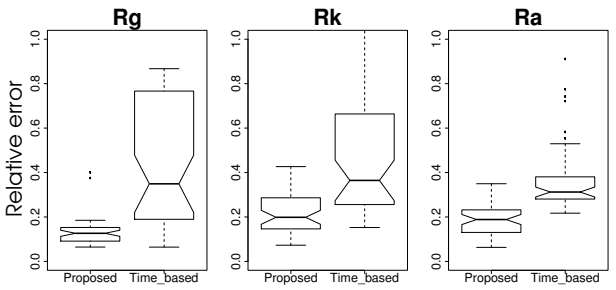

Figure 3: *Distributions of relative error scores for Rg, Rk and Ra, for the proposed method (left box in each panel) and the time-based method (right box), for utterances of 5 male speakers.*

5. Conclusion and future work

This paper presents a new method for automatic voice source parameterisation which attempts to simulate some of the optimisation strategies used when the parameterisation is carried out manually and is guided by both time and frequency domain information. Results showed that voice source parameter data produced by our new method were more similar to the reference data than those of the standard baseline method. Future work will involve incorporating this method into our linguistic studies on the use of the voice source. Furthermore, we hope to apply a similar approach in the development of better automatic inverse filtering.

6. Acknowledgements

This work was supported by the Science Foundation Ireland, Grant 07/CE/I1142 (Centre for Next Generation Localisation, `www.cngl.ie`) and Grant 09/IN.1/I2631 (FASTNET).

7. References

[1] Ní Chasaide, A., Yanushevskaya, I., Gobl, C., "Voice source dynamics in intonation", Proc. of ICPhS, 1470-1473, 2011.

[2] Wong, D., Markel, J., Gray, A., "Least squares glottal inverse filtering from the acoustic speech waveform", IEEE Trans. on Acoustics, Speech and Signal Processing, 27(4), 350-355, 1979.

[3] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", Speech communication, 11, 109-118, 1992.

[4] Drugman, T., Bozkurt, B., Dutoit, T., "Complex cepstrum-based decomposition of speech for glottal source estimation", Proc. of Interspeech, 116-119, 2009.

[5] Gobl, C., Ní Chasaide, A., "Techniques for analysing the voice source," in Coarticulation: Theory, Data and Techniques, W. J. Hardcastle and N. Hewlett, Eds. Cambridge University Press, pp. 300-320, 1999.

[6] Fant, G., Liljencrants, J., Lin, Q. "A four parmeter model of glottal flow", STL-QPSR, 26(4):1-13, 1985.

[7] Fant, G., "The voice source in connected speech", Speech Communication, 125-139, 22(2-3), 1997.

[8] Drugman, T, Dutoit, T., "Glottal closure and opening instant detection from speech signals", in Proc. of Interspeech, 2891-2894, 2009.

[9] Ney, H., "Dynamic programming algorithm for optimal estimation of speech parameter contours", IEEE Transactions on Systems, Man, and Cybernetics 13, 208-214, 1983.

[10] Nelder, J. A., Mead, R., "A simplex method for function minimization", The computer journal, 7(4):308-313, 1965.

[11] Strik, H., Cranen, B., Boves, L., "Fitting a LF-model to inverse filter signals", Proc. of EUROSPEECH-93, 1:103-106, 1993.