Estimation of glottal source waveforms and vocal tract shapes from speech signals based on ARX-LF model

Yongwei Li[1]*, Ken-Ichi Sakakibara*[2]*, Masato Akagi*[1]

[1]Japan Advanced Institute of Science and Technology, Japan
[2] Health Science University of Hokkaido, Japan

yongwei@jaist.ac.jp, kis@hoku-iryo-u.ac.jp, akagi@jaist.ac.jp

Abstract

The widely used method to estimate glottal source waveform and vocal tract shape is to process speech signal using inverse filter and then to fit residual signal using glottal source model. However, since source-tract interactions, estimation accuracy is reduced. In this paper, we propose a method to estimate glottal source waveform and vocal tract shape simultaneously based on analysis-by-synthesis approach with a source-filter model constructed with an auto-regressive eXogenous (ARX) model combined with the Lilijencrant-Fant (LF) model. Since the optimization of multiple parameters makes simultaneous estimation difficult, there are two steps: the glottal source parameters are initialized using the inverse filter method, then the accurate parameters of the glottal source and the vocal tract shape are estimated simultaneously using an analysis-by-synthesis approach. Experimental results with synthetic and real speech signals showed the higher estimation accuracy of the proposed method than inverse filter.

Index Terms: glottal source waveform, vocal tract shape, ARX-LF model

1. Introduction

Estimation of glottal source waveforms and vocal tract shapes is important for speech signal processing. It can be used in many fields, such as speech recognition, speech synthesis, voice pathology detection [1], speaker recognition [2], emotional speech recognition [3], and in further understanding speech production mechanisms. Based on the source-filter theory of speech production, speech signals are modeled as output signals of a vocal tract filter with a glottal source excitation.

There are now many methods for estimating glottal source waveforms and vocal tract shapes based on a source-filter model. A widely used method to estimate vocal tract filters is linear predication (LP) analysis, but the main problem with this method is that it is difficult to estimate vocal tract filters without glottal source effects from speech signals (source-tract interaction) [4]. To overcome this problem, Wong $et\ al.$ estimated glottal source waveforms and vocal tract filters by LP analysis during the glottal closed phase, where there is no interaction [5]. This idea provides reliable estimations only in the long duration of glottal closure. However, it is difficult to find the glottal closed phase in real conditions, especially in the case of a very short glottal closed phase.

A simple and straightforward way to process speech signals to estimate glottal source waveforms is inverse filtering, where glottal sources can be considered residual signals [6, 7]. An improved method was proposed to deal with the residual signals by fitting a Lilijencrant-Fant (LF) model that is one of the widely used glottal source models [8, 9]. The advantage of this method is that a more accurate glottal source model is used, and the disadvantage of this method is source-tract interactions, as mentioned in the above paragraph.

Another method is to estimating glottal source waveforms and vocal tract shapes simultaneously based on an analysis-by-synthesis scheme. The main idea is that a glottal source model is employed as input glottal excitation to a vocal tract filter, and the auto-regressive eXogenous with the LF (ARX-LF) model is used, in which the glottal source signal is represented by the LF model glottal waveform derivative and the vocal tract transfer function is represented by the ARX filter [10]. The advantage of this method is that there is no source-tract interaction because independent glottal sources and vocal tract models were used. However, it is difficult to optimize multiple parameters simultaneously.

To solve this problem, Li $et\ al.$ proposed an iterative algorithm [11] to estimate accurate glottal source waveforms and vocal tract shapes, in which an electro-glotto-graph (EGG) signal was used to estimate initial values for the iteration. It is not convenient to always use EGG.

In this paper, instead of EGG signal, we first obtained the initial values of the LF model parameters using an inverse filter [9]. Then, the accurate glottal source waveforms and vocal tract shapes were estimated simultaneously based on the ARX-LF model using the iterative algorithm [11] .

2. ARX-LF model

According to the source-filter theory of speech production, the glottal source signal in the ARX-LF model is represented by the LF glottal flow derivative and the vocal tract transfer function is represented by the ARX filter. More specifically, the glottal flow derivative is formulated in the LF model by six parameters, where five parameters are related to time T_p, T_e, T_a, T_c and T_0, and one parameter is related to amplitude E_e, as shown in Fig. 1. T_0 is one period of glottal flow, T_p is the instant of the maximum glottal flow model waveform, T_e is the instant of the maximum negative differentiated glottal flow, T_a is the duration of the return phase, T_c is the instant at the complete glottal closure, and E_e the amplitude at the glottal closure instant. T_c is often set to T_0 in a simple LF model version. Thus, five parameters are used in this paper.

A typical LF glottal flow derivative is plotted in Fig. 1. The explicit expression of the LF glottal flow derivative for one fundamental period is given by:

$$u(n) = \begin{cases} E_1 e^{an} \sin(wn), & 0 \le n \le T_e \\ -E_2[e^{-b(n-T_e)} - e^{-b(T_0-T_e)}], & T_e \le n \le T_c \\ 0, & T_c \le n \le T_0 \end{cases} \quad (1)$$

$$s(n) = -\sum_{i=1}^{p} a_i(n) s(n-i) + b_0 u(n) + e(n) \quad (2)$$

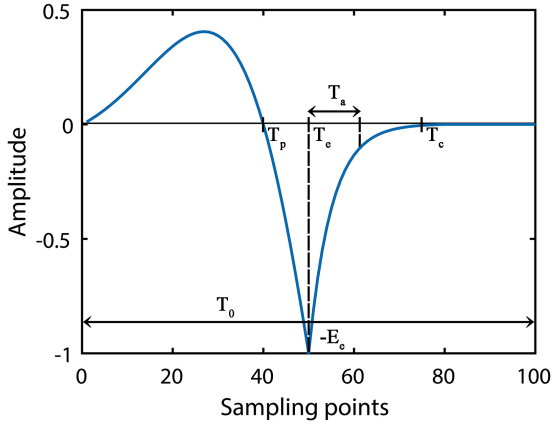

Figure 1: *A typical one period of the LF glottal flow derivative*

where $a_i(n)$ are the coefficients of the p-order ARX model characterizing the vocal tract, b_0 is used to adjust the amplitude of the differentiated glottal flow, and $e(n)$ is the residual signal.

E_1, E_2, a, b and w are the parameters related to T_p, T_e, T_a ,E_e and T_0 [8].

3. Estimation of glottal source waveform and vocal tract shape

The procedure for estimating glottal source waveforms and vocal tract shapes is shown in Fig. 2, and it includes two steps. In the first step, instead of accuracy, initial values are prepared for the next step. The main step is the second step, in which precise glottal source waveforms and vocal tract shapes are estimated simultaneously by the proposed scheme based on the ARX-LF model.

3.1. Initialization

The objective of this step is to determine the period for the LF model. In one period of the LF model waveform, the glottal closure instant (GCI) is a discontinuity location, as shown in Fig. 1, and it is easier to be detected than other locations in one period of the glottal source waveform. Thus, GCI is detected first, and the distance between two continuous GCIs is regarded as one period T_0. The Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) is used for detecting the GCI [12], because it provides more accurate results. The detected GCI by the SEDREAMS method is remarked as GCI_0.

The iterative and adaptive inverse filters（IAIF） and the LF model are used to obtain the initial values of the LF model for simultaneous estimation step. Dynamic programming (DyProg-LF) is employed to fit the LF model parameters, and estimated glottal source parameters (LF model) are remarked as T_p^0, T_e^0, T_a^0 and E_e^0. The detailed implementation of the DyProg-LF algorithm was described in [9, 13].

3.2. Implementation of simultaneous estimation

In this step, a simultaneous estimation algorithm is implemented to accurately estimate a glottal source waveform and vocal tract shape based on the ARX-LF model. There are two processes in this step. In the first process, the LF model parameters and vocal tract filter coefficients can be obtained with a fixed GCI. The initial values of the LF model parameters (T_p^0, T_e^0, T_a^0, E_e^0) are used for synthesizing glottal source waveform derivative $u(n)$, and $u(u)$ is then exploited to synthesize $x(n)$ using the ARX model, and the ARX model parameters can be estimated with mean square error (MSE) sense for $e(n)$ by the least square (LS) method. In each iteration of this optimization process, the LF model parameters are regenerated around the initial values of T_p^0, T_e^0, T_a^0 and E_e^0, and a glottal source waveform derivative can be re-generated using these parameters. Let a vector β:

$$\beta(n) = [-s(n-1)\cdots - s(n-p)u(n)]^{\mathrm{T}} \tag{3}$$

and the ARX model coefficients γ are represented as follows:

$$\gamma = [a_1 \cdots a_p b_0]^{\mathrm{T}} \tag{4}$$

γ can be calculated as follows:

$$\gamma = [\beta(n)\beta(n)^{\mathrm{T}}]^1\beta(n)s(n) \tag{5}$$

In the second process, we can estimate more accurate LF model parameters and vocal tract coefficients. Since the performance of the ARX-LF model is affected by the accuracy of GCI which was reported by Lu [14], we suggested shifting the parameters of GCI around the value of GCI_0 further. Then, the first process works again for each shifted GCI. For the given GCI each time, the iteration processing in the minimization of mean square error (MMSE) optimization is set to 2000. The accurate glottal source parameters and vocal tract filter coefficients are finally estimated by MMSE.

The sampling frequency was 12000 Hz and p was set to 14 in this paper. The estimation length of the frame is three periods of glottal source waveforms, and the shift frame is one period of a glottal source waveform.

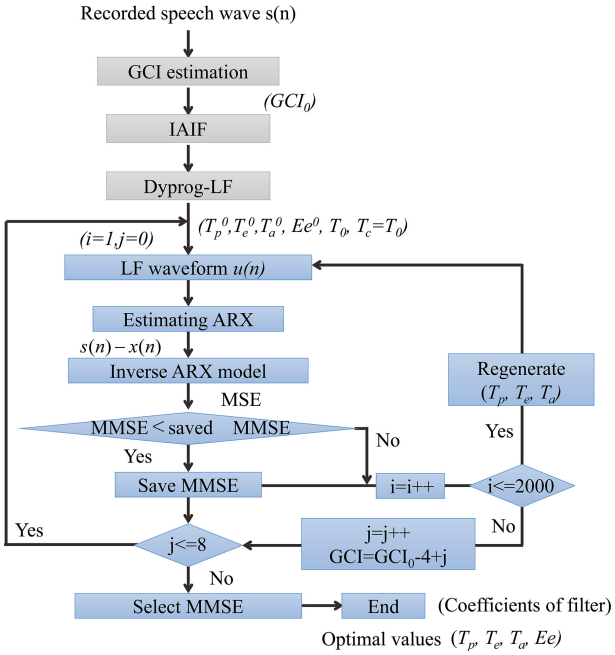

Figure 2: *Estimation scheme of glottal source waveform and vocal tract shape*

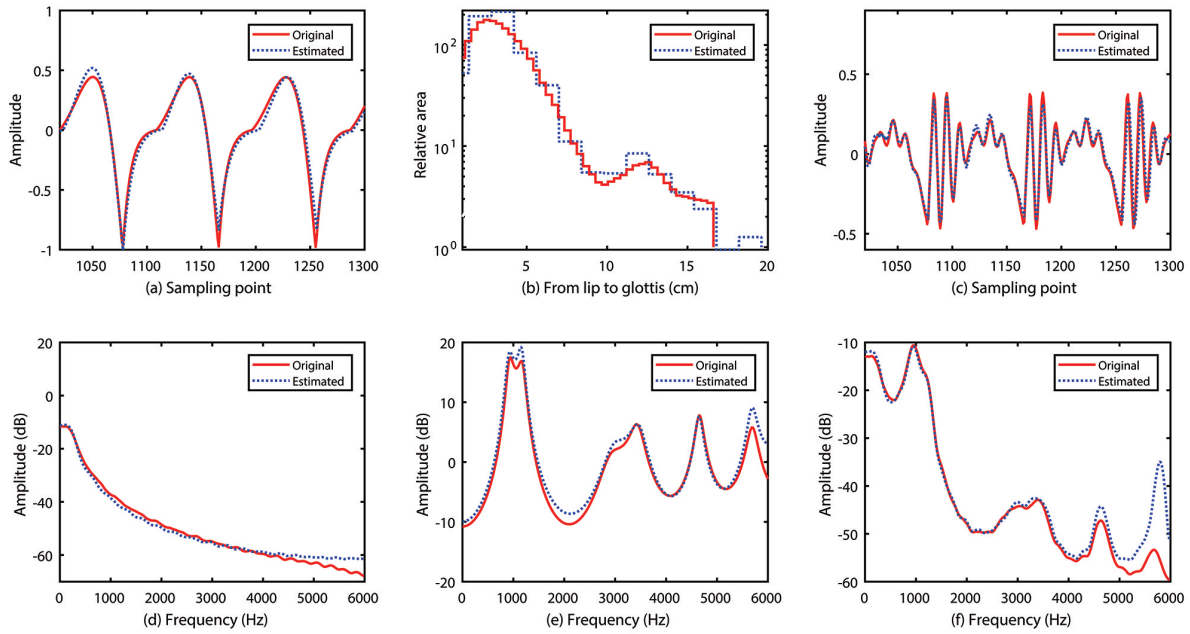

Figure 3: *Original glottal source waveform and estimated glottal source waveform in time domain (a) and frequency domain (d), original vocal tract shape and estimated vocal tract shape in time domain (b) and frequency domain (e), original voice waveform and estimated voice waveform in time domain (c) and frequency domain (f).*

Table 1: *Average estimation errors (ε) for synthesized vowels of two methods*

	Glottal source				**Vocal tract**	
	$T_p(\%)$	$T_e(\%)$	$T_a(\%)$	$E_e(\%)$	$F_1(\%)$	$F_2(\%)$
IAIF-Dyprog-LF	24.0	19.8	50.4	82.2	9.4	6.1
Proposed	11.4	9.5	60.0	23.2	2.3	1.1

4. Experimental results

First, the synthetic vowels were used to test the proposed estimation method and the IAIF with Dyprog-LF method in [9]. The advantage of testing on synthetic vowels is that the accuracy of the proposed method can be investigated by comparing the estimated parameter values of glottal source waveforms and vocal tract shapes with referenced parameter values. Then, the glottal source waveforms and vocal tract shapes of real vowels are estimated to test the proposed estimation method and the IAIF with Dyprog-LF.

4.1. Synthesized vowels

The synthesized vowels are produced according to the source-filter model, in which a glottal source waveform derivative is generated by the LF model, and the vocal tract filters are taken from five Japanese vowels (/a/, /e/, /i/, /o/ and /u/) using Kawahara's method [15]. This is because of formant frequencies of vocal tract in these vowels vary widely, especially the first and second important formant frequencies (F_1 and F_2). A larger number of synthetic vowels with a wide range of LF model parameter values are used in this paper. The LF model parameters are varied: T_e: 0.3 to 0.9 with steps of 0.05; T_p/T_e: 0.65 to 0.8 with steps of 0.05 steps; T_a: 0.03, 0.08, within a suggested range in [16]. In order to synthesize more realistic vowels, the fundamental frequency (F_0) is obtained from a real vowel, 18 different F_0 ranged from 90 to 140 Hz are used for synthesizing vowels. The total number of synthesized vowels with 9360 ($4[T_p] \times 13[T_e] \times 2[T_a] \times 18[F_0] \times 5[filter]$) different conditions are used for testing proposed method.

4.1.1. Results and discussion

Estimated values of the LF model parameters and F_1 and F_2 of the vocal tract were evaluated by the reference values Let the reference values as a vector $\theta \in \{T_p, T_e, T_a, E_e, F_1, F_2\}$ and the estimated values as vector $\hat{\theta}$. The error (ε) between estimation and reference values can be calculated as:

$$\varepsilon = \frac{|\hat{\theta} - \theta|}{\theta} \times 100\% \quad (6)$$

The average errors (ε) are listed in Table 1. Estimation errors of a glottal source are less than 12 except for those of T_a, since T_a was the smallest of all parameters as the denominator in Eq. 6 and the error was 60%. Estimation errors are less than 2.3% for the vocal tract. Figure 3 shows an example of estimated results, in which the glottal source waveform and vocal tract shape are estimated from a synthesized vowel /a/. As shown in Fig. 3, estimated glottal source waveforms and vocal tract shapes are very similar to the original ones in the time

Table 2: *Average estimation errors (ε_{OQ}), and F_1 and F_2 are estimated by the ARX model and Praat from five males and five females*

Male (M) and Female (F)	M1	M2	M3	M4	M5	F1	F2	F3	F4	F5
$\varepsilon_{OQ}(IAIF-Dyprog-LF)$ [%]	9.0	15.3	2.0	18.8	11.1	10.8	12.6	8.9	0.2	10.3
$\varepsilon_{OQ}(Proposed)$ [%]	12.6	13.0	1.5	0.2	7.1	4.9	1.5	2.7	9.0	10.8
$F_{1\ ARX}$ (Hz)	773	756	680	803	709	1031	1031	797	1125	1043
$F_{1\ Praat}$ (Hz)	734	740	675	788	686	1064	997	722	1098	1023
$F_{2\ ARX}$ (Hz)	1348	1189	1213	1313	1137	1611	1553	1025	1605	1459
$F_{2\ Praat}$ (Hz)	1334	1194	1213	1315	1129	1587	1506	1013	1587	1464

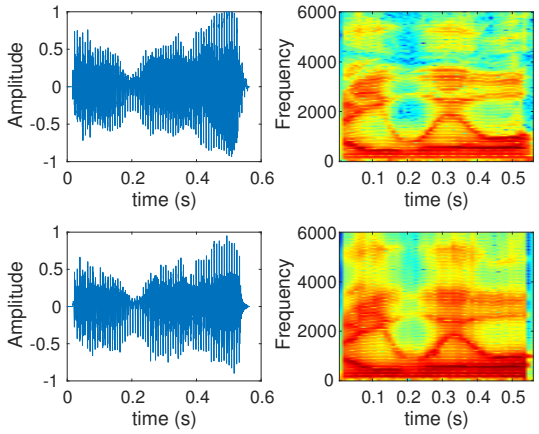

Figure 4: *Original speech waveform and its spectrogram (top), re-synthesized speech waveform and its spectrogram (bottom)*

domain and frequency domain. The length of the vocal tract shapes are different between estimated and original one because the sampling frequency and the order of the vocal filters is different between the synthesis step (Kawahara's method: 44100-Hz sampling frequency with an order 44th order) and the analysis step (ARX-LF: 12000-Hz sampling frequency with 14th order).

Table1 shows that the estimation accuracy of the proposed method is higher than that of IAIF with Dyprog-LF.

4.2. Natural vowels

The voiced vowel (/a/) was pronounced by five male and five female Japanese speakers. The speech signals were recorded together with electroglottographic (EGG) signals. Thus, there are ten real voiced vowels used to test the proposed method and the IAIF with Dyprog-LF.

4.2.1. Results and discussion

There is no direct reference parameter available for the glottal sources and vocal tracts in real vowels. To evaluate glottal sources, as a reference value, we calculated the open quotient (OQ) to evaluate the accuracy of the proposed method. The recorded vowels were analyzed by the proposed method to estimate T_e, which is often considered as the OQ_{LF}, and referenced OQ_{EGG} was calculated from a differentiated EGG (dEGG) signal by searching glottal opening instant (GOI) and GCI. Thus, the estimation errors can be calculated by Eq. 6. The estimation errors (ε) are listed in Table 2. Compared with the value of OQ obtained from the dEGG signal, the accuracy of the proposed method is higher than that of IAIF with Dyprog-LF.

Vocal tract parameters F_1 and F_2 were estimated by the proposed method and a widely used formant extractor (Praat), respectively. Results are shown in Table 2, where the values of F_1 and F_2 estimated by the proposed method are very similar to those extracted by Praat. Furthermore, for ten speakers, most of the values of F_1 estimated by proposed method are a little higher than those estimated by Prrat, and the values of F_2 of two the methods are mostly the same. Therefore, the proposed method can effectively estimate the vocal tract parameters.

Moreover, a continuous speech (/aiueo/) pronounced by a male speaker was challenged by the proposed method. It is impossible to discuss glottal source parameters since there was no EGG signal recorded together with a speech signal. The waveform and the spectrogram of the original speech signal and the resynthesized speech signals by the ARX-LF model are plotted in Fig. 4, which shows that the original speech signal is very similar to the resynthesized speech signal in the time and frequency domain. The spectrogram clearly shows that the formant frequencies are the same as the original one, which proves the high accuracy of the proposed method in estimating vocal tracts of continuous speech. The synthesized speech can be perceived as well as original one by an informal perception test. Therefore, the proposed method is also suitable for continuous speech. The slight difference between the original speech and the resynthesized one may be caused by using only the ARX-LF model, in which $e(n)$ was not added in the synthesis process.

All of results demonstrate that the proposed method has higher estimation accuracy than that of IAIF with Dyprog-LF, and proposed method is available for continuous speech.

5. Conclusion

In this paper, we proposed a simultaneous estimation of glottal source waveforms and vocal tract shapes from speech signals based on the ARX-LF model. The estimation procedures contain two steps: obtaining the initial values of glottal source parameters, in which an inverse filter and the LF model are used, and using a simultaneous estimation procedure to obtain accurate glottal sources and vocal tract parameters with the ARX-LF model. We tested both the synthesized vowels and real vowels with the proposed method and IAIF with Dyprog-LF method. The results show that the proposed method has higher estimation accuracy than that of IAIF with Dyprog-LF. Moreover, the proposed method shows robustness for continuous speech.

6. Acknowledgements

This study was supported by a Grant-in-Aid for Scientific Research (A) (No. 25240026) and China Scholarship Council (CSC).

7. References

[1] T. Drugman, T. Dubuisson, and T. Dutoit, "On the mutual information between source and filter contributions for voice pathology detection," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[2] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, 1999.

[3] A. I. Iliev, M. S. Scordilis, J. P. Papa, and A. X. Falcão, "Spoken emotion recognition through optimum-path forest classification using glottal features," *Computer Speech & Language*, vol. 24, no. 3, pp. 445–460, 2010.

[4] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice-hall Englewood Cliffs, NJ, 1978, vol. 100.

[5] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[6] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[7] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," 2009.

[8] G. Fant, J. Liljencrants, and Q.-g. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[9] J. Kane and C. Gobl, "Automating manual user strategies for precise voice source analysis," *Speech Communication*, vol. 55, no. 3, pp. 397–414, 2013.

[10] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of lf glottal source parameters based on an arx model," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[11] Y. Li, K.-I. Sakakibara, D. Morikawa, and M. Akagi, "Commonalities of glottal sources and vocal tract shapes among speakers in emotional speech," in *International Seminar on Speech Production (ISSP)*, Tianjing, China, 2017, pp. 79–81.

[12] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2012.

[13] J. Kane, I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "Exploiting time and frequency domain measures for precise voice source parameterisation," in *Speech Prosody 2012*, 2012.

[14] H.-L. Lu, *Toward a high-quality singing synthesizer with vocal texture control*. Stanford University, 2002.

[15] H. Kawahara, K.-I. Sakakibara, H. Banno, M. Morise, T. Toda, and T. Irino, "Aliasing-free implementation of discrete-time glottal source models and their applications to speech synthesis and f0 extractor evaluation," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*. IEEE, 2015, pp. 520–529.

[16] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech & Language*, vol. 26, no. 1, pp. 20–34, 2012.