INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Iterative Optimal Preemphasis for Improved Glottal-Flow Estimation by Iterative Adaptive Inverse Filtering

Parham Mokhtari and Hiroshi Ando

Center for Information and Neural Networks (CiNet)
National Institute of Information and Communications Technology (NICT), Japan
parham@nict.go.jp

Abstract

Iterative adaptive inverse filtering (IAIF) [1] remains among the state-of-the-art algorithms for estimating *glottal flow* from the recorded speech signal. Here, we re-examine IAIF in light of its foundational, classical model of voiced (non-nasalized) speech, wherein the overall spectral tilt is caused only by lip-radiation and glottal effects, while the vocal-tract transfer function contains formant peaks but is otherwise not tilted. In contrast, IAIF initially models and cancels the formants after only a first-order preemphasis of the speech signal, which is generally not enough to completely remove spectral tilt.

Iterative optimal preemphasis (IOP) is therefore proposed to replace IAIF's initial step. IOP is a rapidly converging algorithm that models a signal (then inverse-filters it) with one real pole (zero) at a time, until spectral tilt is flattened. IOP-IAIF is evaluated on sustained /a/ in a range of voice qualities from weak-breathy to shouted-tense. Compared with standard IAIF, IOP-IAIF yields: (i) an acceptable glottal flow even for a weak breathy voice that the standard algorithm failed to handle; (ii) generally smoother glottal flows that nevertheless retain pulse shape and closed phase; and (iii) enhanced separation of voice qualities in both normalized amplitude quotient (NAQ) and glottal harmonic spectra.

Index Terms: speech analysis, glottal inverse filtering, optimal preemphasis

1. Introduction

Glottal inverse filtering refers to the process of estimating the source of voiced speech sounds (the glottal volume-velocity waveform, known as *glottal flow*), most conveniently and non-invasively from the acoustic pressure signal recorded by a microphone. The approach common to most algorithms proposed over the past six decades [2], is to remove the acoustic effects of the supralaryngeal system by some kind of modeling followed by filtering of the speech signal through the inverse of the model, thus leaving only the laryngeal (excitation or source) signal.

However, while several techniques exist for measuring various aspects of glottal kinematics, glottal flow itself has never been directly measured, and so there is no absolute ground-truth data with which to compare the results of inverse filtering algorithms. Despite this seemingly dire situation, researchers have a basic conception of what types of estimated glottal flow reasonably conform (or not) with expectations regarding the physics of vocal-fold oscillation; e.g., the two most common criteria are the relative absence of time-domain ripples that are caused by incomplete cancelation of formants, and the presence of a relatively flat portion of the waveform during the closed phase. Meanwhile, ongoing research tries to improve quantification of the quality of estimated flow [3] and to automate experts' subjective choices [4].

One of the most widely used algorithms for glottal inverse filtering, and one that is still regarded as an important benchmark [5], is iterative adaptive inverse filtering (IAIF) [1]. It is founded on the classical model of speech production as a linear cascade of three processes (at least for non-nasalized voiced sounds) [6]: G, the glottal source which provides the volume-velocity excitation; V, the vocal-tract airway that imparts resonances which appear as formant peaks in the spectrum; and L, the lip-radiation effect which is essentially a differentiator converting volume-velocity at the lips to farfield acoustic pressure. Thus in the z-domain, the recorded speech sound S is written as:

$$S(z) = G(z)V(z)L(z)\,, \tag{1}$$

where lip-radiation is of the form:

$$L(z) = 1 - bz^{-1}, \quad 0 < b \le 1\,. \tag{2}$$

Assuming fixed values for lip-radiation coefficient b and vocal-tract autoregressive model order M, IAIF is an automatic algorithm that, in two main iterations, tries to model (by linear prediction analysis) and cancel (by inverse filtering) V and L, in order to leave only G. In practice, to get the best estimate of glottal flow one usually runs IAIF for a range of values of b and M, then chooses the result that best satisfies certain criteria as mentioned earlier (whether subjective as is most often the case, or partly automated as, e.g., in [7]).

A fundamental property of the separated speech model is that the *overall spectral tilt* of the recorded pressure signal is a combination of a downward tilt imparted by G (the degree of tilt dependent on laryngeal voice quality) and an upward tilt imparted by L (dependent on lip aperture), with *no explicit contribution by V to overall spectral tilt*. Indeed, as evidenced in key literature, e.g. [6, Fig. 1.3-1] [8, Fig. 3.23], the volume-velocity transfer function of an idealized V does not have an overall tilt, other than that which occurs as a result of a non-uniform distribution of formant frequencies and/or bandwidths. The classical view is further supported by more recent acoustic measurements and simulations using physical vocal-tract models, e.g. [9, Fig. 4] [10, Fig. 5].

To be consistent with this view, modeling of V ought to be performed only after having removed any existing spectral tilt. However, prior to modeling of V, as a first step IAIF tries to approximately cancel the combined effects of G and L by only 1st-order autoregressive modeling and inverse filtering. A 2nd- or higher-order model is wisely not used in the first step, because that would unintentionally remove one or more of the

http://dx.doi.org/10.21437/Interspeech.2017-79

vocal-tract resonances; a 1st-order model is used to ensure a single, real pole at DC, which represents mainly spectral tilt. Nevertheless, in light of the classical assumptions, the question arises whether a 1st-order model is sufficient in general to achieve a flat-tilted, V-only transfer function. To achieve better consistency with the classical model, here we propose to replace IAIF's first step with *iterative optimal preemphasis* which, as explained in section 3, guarantees a signal with flat spectral tilt over the available frequency range.

2. Speech Data & Preprocessing

The speech material recorded for this study was /a/ sustained for about 1.5 s by an adult, male speaker, in each of five voice qualities along a continuum from weak & breathy, to loud & tense: (i) weak and breathy voice, (ii) breathy voice, (iii) modal voice, (iv) loud and slightly tense voice, and (v) shouted and tense voice. The vowel /a/ was chosen as it is usually a good candidate for glottal inverse filtering, due to its first formant being separated from the fundamental frequency.

For high signal-to-noise ratio and low amplitude and phase distortions, the speech data were recorded in a sound-treated room, with a condensor microphone (B&K 4190) at a constant distance of about 25 cm from the speaker's lips; a conditioning amplifier (B&K Nexus) set to unity gain; and an audio interface (RME Babyface Pro) connected to a laptop PC, with 44.1 kHz sampling rate and 24 bits/sample. (To hear the recordings, please refer to the accompanying MP3 audio files.)

Prior to inverse filtering, the recorded signal's polarity was corrected by negation, the signal was downsampled to a lower rate of 8 kHz [11], and a linear phase high-pass filter with a 70 Hz cut-off was applied to suppress any low-frequency ambient disturbances. One representative frame of duration 50 ms was located one-third of the way into the voiced segment of each sustained /a/.

Each /a/ token was produced at a fundamental frequency that was natural and comfortable for the corresponding loudness and voice quality (cf. Table 1). Also listed in Table 1 is the signal energy in each analysis frame: as expected, calculated energy increased monotonically from the softest to the loudest utterance, with an overall dynamic range of 29 dB.

3. Iterative Optimal Preemphasis (IOP)

Optimal preemphasis (OP) is often used as an initial step in speech processing, motivated by the fact that, in reducing overall tilt, it tends to reduce the spectral dynamic range and improve the stability of subsequent modeling and feature extraction [12, p. 216] [13, p. 574]. However, the spectral tilt of voiced sounds depends not only on the individual speaker and gender, but also varies considerably with vocal effort and laryngeal voice quality [14-16]. Therefore, OP can certainly reduce, but rarely eliminate, spectral tilt.

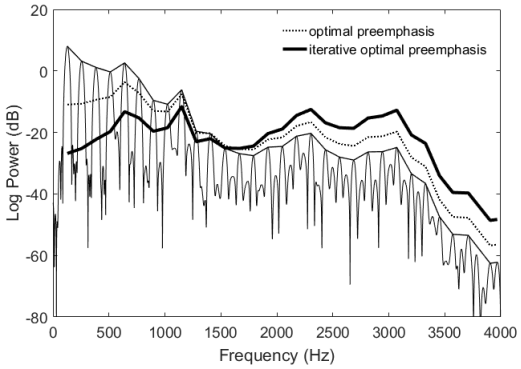

Figure 1. *Effects of conventional OP and proposed IOP, on harmonic spectrum of /a/ (modal voice).*

As is well known, OP involves modeling the signal with 1st-order linear prediction (LP) analysis (i.e., a single pole at $a_1 = R_1 / R_0$ on the real axis in the z-plane, where R_n is the autocorrelation at lag n samples), then filtering the signal with the inverse of the obtained model ($1 - a_1 z^{-1}$, which imposes a zero at the same location in the z-plane) [17, p.215]. However, even the most aggressive preemphasis with $a_1 = 1.0$ is only able to modify the spectral tilt by up to 6 dB/oct, which may not be enough to achieve zero tilt. This view is supported by the simple observation that, after applying OP to a frame of speech, a new value for a_1 computed from the preemphasized signal will generally not equal zero; this implies, as mentioned above, that only one application of OP is generally insufficient to remove spectral tilt.

We have found that repeated application of OP, here termed *iterative* optimal preemphasis (IOP), yields a monotonically decreasing sequence of $|a_1|$ that converges towards 0. In practice, it is useful to stop the iterations by setting a threshold (e.g., as soon as $|a_1| < 0.001$), and the resulting signal can then be considered to have a flat overall spectral tilt. More precisely, in contrast to OP, IOP completely removes the speech signal's autocorrelation at lag 1 sample.

Fig. 1 shows an example of the log-power spectrum of /a/ with normal (modal) phonation, and the contrast between conventional OP and IOP. The thin lines joining the spectral peaks depict the outline of the measured harmonic spectrum. The dotted lines display the harmonic spectrum after OP, with $a_1 = 0.944$; evidently, there remains a downward spectral tilt. In contrast, the thick lines show the harmonic spectrum after IOP, which in this case converged in just 6 iterations with the following sequence of a_1 values: 0.944, 0.749, 0.341, 0.070, 0.016, and 0.004; the next value for $|a_1|$ was less than 0.001, implying that in terms of real-pole LP modeling the log-power spectrum was optimally flat-tilted.

One of many possible methods of quantifying spectral tilt, is to perform linear regression on the spectral harmonics on a

Table 1. *Fundamental frequency of voicing, energy, and a_1 sequence yielded by IOP, for each of the 5 recorded voice qualities.*

Voice Quality	F0 (Hz)	Energy (dB re:Modal)	preemphasis coefficients a_1 yielded by IOP (bold font: OP)
Shouted & Tense	161	19.6	**0.751**, 0.446, 0.128, 0.037, 0.012, 0.004, 0.001
Loud	161	13.7	**0.823**, 0.577, 0.220, 0.064, 0.021, 0.007, 0.002
Modal	128	0	**0.944**, 0.749, 0.341, 0.070, 0.016, 0.004
Breathy	116	−6.1	**0.991**, 0.858, 0.121, 0.008
Weak & Breathy	134	−9.4	**0.993**, 0.898, 0.248, 0.047, 0.008, 0.001

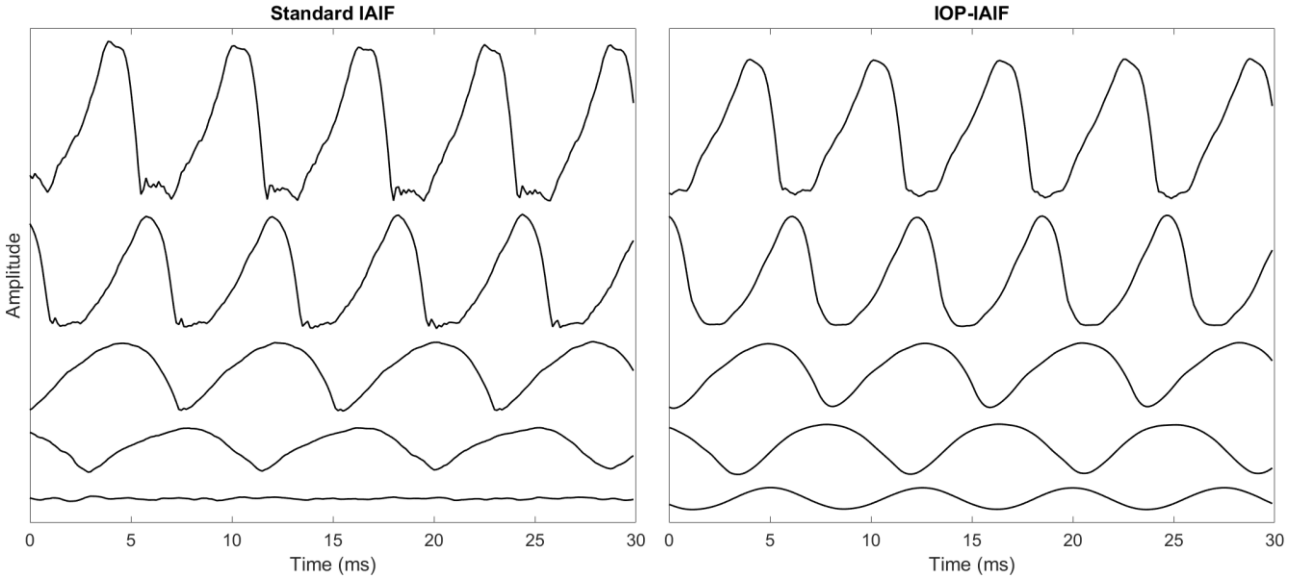

Figure 2. *Glottal flow waveforms estimated by standard- vs IOP-IAIF. Amplitude scales are identical in both panels; waveforms were shifted in amplitude for visual clarity. From top to bottom: shouted & tense, loud, modal, breathy, weak & breathy.*

log frequency scale. Such analysis on the spectra in Fig. 1 yielded −13.1 dB/oct for the original harmonics, −8.0 dB/oct after OP (a reduction by only 5.1 dB/oct, as expected), and −2.0 dB/oct after IOP (this value is not exactly 0, because as mentioned, there are many ways of measuring tilt). These numbers confirm that IOP reduces spectral tilt more effectively than conventional OP.

The sequence of a_1 values yielded by IOP for each voice quality is listed on the right side of Table 1. The first value (in bold font) matches a_1 for conventional OP. By comparison, IOP rapidly converged within 4 to 7 iterations. Interestingly, both the first and second values of a_1 vary monotonically with voice quality, indicating stronger preemphasis required for weak & breathy voice and weaker preemphasis for shouted & tense voice, as expected.

It is worth noting that, just as IAIF may use either conventional LP analysis or more sophisticated, discrete all-pole modeling (DAP) [18], IOP can also operate either directly on a frame of speech signal (with R_n calculated by time-domain autocorrelation) or on a discrete, harmonic power spectrum (with R_n calculated by discrete Fourier transformation). This study uses LP modeling and time-domain autocorrelation; the potential for discrete spectral modeling to yield greater accuracy is left for future work.

4. Preliminary evaluation of IOP-IAIF

We now compare the performance of standard- and IOP-IAIF (i.e., before / after replacing 1st-order LP modeling with IOP), in glottal inverse filtering of the five recordings. In all cases, the order of LP analysis for glottal flow modeling was set to 4.

To allow meaningful comparison of estimated glottal flow amplitudes across different vocal efforts or voice qualities, we ensured a constant mouth-to-microphone distance, and our implementation of IAIF included level adjustment of the vocal-tract LP model to enforce unity gain at DC [19]. Indeed, from the weakest to the loudest vocalization, Fig. 2 shows that the estimated glottal flows increase monotonically in both peak-to-peak amplitude and steepest negative slope — two parameters that are known to be related with the amplitude of the fundamental and the speech sound pressure level, respectively [20].

Independently for each analysis we varied the vocal-tract model order M (from 8 to 18, in steps of 2) and lip-radiation coefficient b (from 0.80 to 0.99, in steps of 0.01) in search of the best result, i.e., a glottal flow best satisfying the two subjective criteria stated in section 1, and a vocal-tract model spectrum with no spurious (non-formant) peaks. This proved successful in 8 cases out of 10; the two problematic cases were standard IAIF analyses of breathy, and weak & breathy voice.

For breathy voice, the glottal flow estimated by standard IAIF (second from the bottom in Fig. 2) appeared to be not unreasonable; but the final vocal-tract LP model (cf. Fig. 3a) showed a spurious peak far below the first formant, in the vicinity of the first two harmonics.

More critically, as shown in the bottom left waveform of Fig. 2, standard IAIF did not offer any reasonable glottal flow

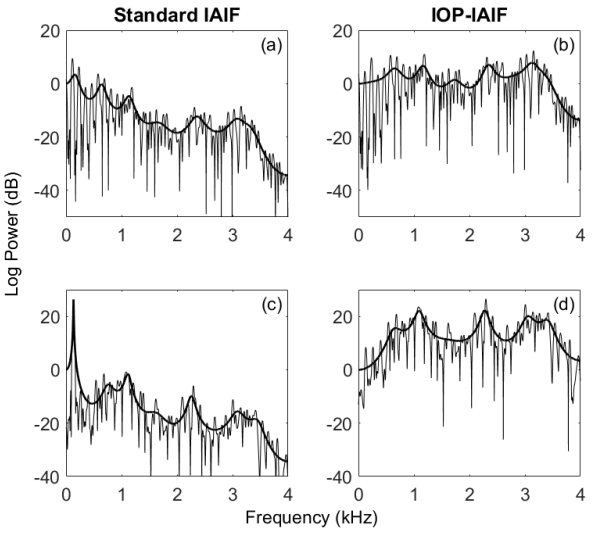

Figure 3. *Vocal-tract spectra and LP models for breathy voice (a & b) and weak & breathy voice (c & d).*

for weak & breathy voice. The reason for this is shown in Fig. 3c: the vocal-tract LP model erroneously included a sharp peak at the fundamental, as it clearly dominated the spectrum. In contrast, Fig. 3d shows that IOP not only flattened the spectral tilt but also effectively suppressed the fundamental, resulting in a more reasonable modeling of only the formants. Consequently, IOP-IAIF yielded a nearly sinusoidal glottal flow (cf. bottom right of Fig. 2), which is characteristic of a fundamental-dominated, weak & breathy voice.

Moving up in Fig. 2 to modal, loud, and shouted voices, it is clear that IOP-IAIF yielded consistently smoother waveforms, almost free of the jagged, noisy behavior seen especially in the closed phase of glottal flow yielded by standard IAIF. It is important to note that this property of IOP-IAIF glottal flows is not the same as merely smoothing the standard-IAIF waveforms; rather, owing to the flattening of spectral tilt, IOP-IAIF yields a vocal-tract model that is at least subtly, and as shown in Fig. 3 sometimes radically, different compared with the corresponding model in standard IAIF. Therefore, IOP-IAIF glottal flow appears smoother not only due to spectral tilt, but also due to the fine balance among harmonic amplitudes thanks to improved vocal-tract modeling.

The waveforms in Fig. 2 also indicate that while the IOP-IAIF glottal flows are smoother, they retain important features such as the breathy-to-tense tendency towards smaller open-quotients and faster closing-speeds. The normalized amplitude quotient (NAQ) is a well-known parameter that has been shown to be related to voice quality variations along the breathy-to-tense continuum [21]. Fig. 4 shows the mean of NAQs extracted from the central 3 to 5 periods of each estimated glottal flow (NAQ for weak & breathy voice analyzed by standard IAIF is not included because the glottal flow in this case was simply meaningless). In line with the literature, the values in Fig. 4 indicate that breathy voice has higher NAQ while tense voice has lower NAQ, with modal voice in between. While NAQ calculated from IOP-IAIF glottal flows are consistently slightly higher (towards the breathy side) compared with standard IAIF, Fig. 4 also shows that they are better separated among the voice qualities: even excluding weak & breathy voice, the range (and ratio) of NAQ for breathy versus shouted was 0.10 (1.98) for standard IAIF, and 0.16 (2.32) for IOP-IAIF. Moreover, among the 3-5 glottal pulses analyzed in each case, the standard deviation of NAQ was on average 0.016 for standard IAIF, and only 0.005 for IOP-IAIF; this suggests that IOP-IAIF can provide greater consistency in the shape of consecutive glottal pulses estimated within one analysis frame.

Finally, Fig. 5 compares the glottal harmonic spectra of the five voice qualities across analysis conditions. To

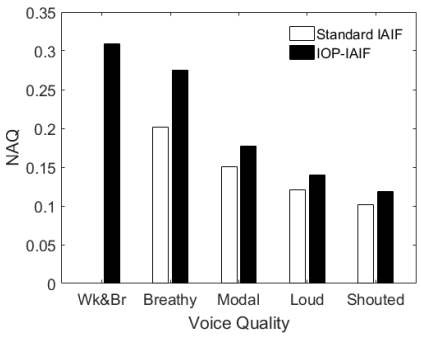

Figure 4. *Mean values of normalized amplitude quotient (NAQ) in 3-5 central periods of estimated glottal flow.*

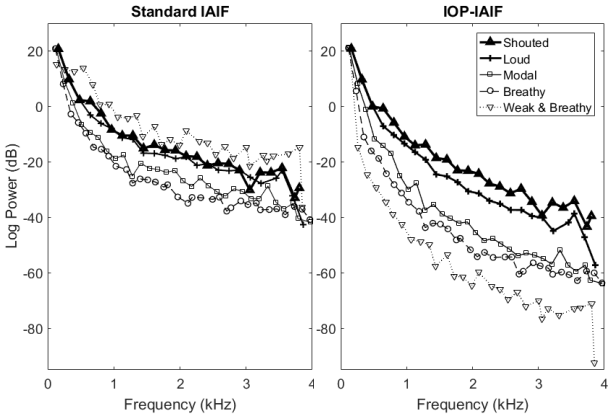

Figure 5. *Energy-normalized harmonic spectral envelopes of estimated glottal flow.*

emphasize the effects of voice quality on spectral balance rather than overall sound pressure level, each spectrum (on which the harmonics were measured) was normalized to an energy of 0 dB. Thanks to optimal flattening of vocal-tract spectral tilt, IOP-IAIF allocated a steeper tilt to glottal spectra; hence the smoother glottal flows in Fig. 2. Furthermore, even ignoring standard-IAIF's erroneous result for weak & breathy voice, the IOP-IAIF glottal harmonic spectra clearly show a better separation of the five voice qualities — i.e., a greater dynamic range at most frequencies; and a clear progression of the glottal component of spectral tilt, from weak & breathy (most tilted) to shouted & tense (least tilted).

5. Conclusions

This study proposed iterative optimal preemphasis (IOP) as a replacement for 1st-order LP modeling in the first step of the IAIF glottal inverse filtering algorithm. The motivations for this are grounded in the acoustic theory of speech production, wherein the ideal transfer function of the vocal-tract, separated from glottal and lip-radiation effects, has no overall tilt. IOP is a rapidly converging algorithm that removes the speech signal's autocorrelation at lag 1 sample, thereby removing spectral tilt across the available frequency range.

The proposed method (IOP-IAIF) was evaluated by comparing its performance with standard IAIF, on a small dataset of sustained /a/ in five distinct voice qualities. In the time domain, estimated glottal flows were smoother, while retaining pulse shape, skewness, and the relatively flat portions of the closed phase. Moreover, the NAQ parameter calculated from the estimated flow signals retained a monotonic relation with voice quality and an expanded range from tense to breathy voice. In the frequency domain, the glottal harmonic spectra also showed a wider dynamic range and better separation among the voice qualities.

Although the evaluation here was limited in terms of the size and scope of the speech data, these preliminary yet in-depth results are promising. As both standard- and IOP-IAIF still require a human expert to judge and select the best results, more extensive evaluations with larger datasets including speakers of both genders, different vowels, and a greater variety of voice qualities and fundamental frequencies, will be important but labor-intensive. We hope that this study can stimulate further research on improving and automating glottal inverse filtering algorithms, for both basic and applied studies of the human voice.

6. References

[1] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication*, vol. 11, pp. 109–118, 1992.

[2] P. Alku, "Glottal inverse filtering analysis of human voice production — A review of estimation and parameterization methods of the glottal excitation and their applications," *Sādhanā*, vol. 36, part 5, pp. 623–650, 2011.

[3] E. Moore and J. Torres, "A performance assessment of objective measures for evaluating the quality of glottal waveform estimates," *Speech Communication*, vol. 50, pp. 56-66, 2008.

[4] J. Kane and C. Gobl, "Automating manual user strategies for precise voice source analysis," *Speech Communication*, vol. 55, pp. 397-414, 2013.

[5] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech and Language*, vol. 26, pp. 20-34, 2012.

[6] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, (2nd printing) 1970.

[7] M. Airaksinen, T. Bäckström, and P. Alku, "Automatic estimation of the lip radiation effect in glottal inverse filtering," in *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association, September 14–18, Singapore, Proceedings*, 2014, pp. 398–402.

[8] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. Berlin, Heidelberg, New York: Springer, (2nd edition) 1972.

[9] T. Kitamura, H. Takemoto, S. Adachi, and K. Honda, "Transfer functions of solid vocal-tract models constructed from ATR MRI database of Japanese vowel production," *Acoustical Science & Technology*, vol. 30, no. 4, pp. 288–296, 2009.

[10] H. Takemoto, P. Mokhtari, and T. Kitamura, "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method," *J. Acoustical Society of America*, vol. 128, no. 6, pp. 3724–3738, 2010.

[11] P. Alku and E. Vilkman, "Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering," *J. Acoustical Society of America*, vol. 98, no. 2, pp. 763–767, 1995.

[12] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Berlin, Heidelberg, New York: Springer, 1976.

[13] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[14] A. M. C. Sluijter and V. J. van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *J. Acoustical Society of America*, vol. 100, no. 4, pp. 2471–2485, 1996.

[15] G. Fant, "The voice source in connected speech," *Speech Communication*, vol. 22, pp. 125-139, 1997.

[16] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, pp. 189-212, 2003.

[17] A. H. Gray, Jr. and J. D. Markel, "A Spectral-Flatness Measure for Studying the Autocorrelation Method of Linear Prediction of Speech Analysis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 22, no. 3, pp. 207–217, 1974.

[18] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[19] P. Alku, E. Vilkman, and A.-M. Laukkanen, "Estimation of amplitude features of the glottal flow by inverse filtering speech pressure signals," *Speech Communication*, vol. 24, pp. 123-132, 1998.

[20] J. Gauffin and J. Sundberg, "Spectral correlates of glottal voice source waveform characteristics," *J. Speech, Language, and Hearing Research*, vol. 32, pp. 556-565, 1989.

[21] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.