INTERSPEECH 2013

Periodicity extraction for voiced sounds with multiple periodicity

Masanori Morise[1], *Hideki Kawahara*[2], *Kenji Ozawa*[1]

[1]Faculty of Engineering, University of Yamanashi, Japan
[2]Department of Design Information Sciences, Wakayama University, Japan
mmorise@yamanashi.ac.jp

Abstract

A periodicity extraction method is introduced to analyze voiced sounds with a complex excitation behavior. Although general voiced sound has only one periodicity, some voiced sounds such as the pathological voice and the singing voice often have multiple periodicities. A method for estimating multiple periodicities from voiced sounds to deal with these kinds of voices is proposed in this article. At first, a definition of the multiple periodicity and its causes are explained, and then the principle of the proposed method is introduced. The proposed method was evaluated by using several artificial signals and pathological voices recorded in a real environment. The analysis results from the artificial signals indicated that the proposed method can extract multiple periodicities, and that of the pathological voices shows a similar tendency. These results suggest that the proposed method is effective at extracting the multiple periodicities.

Index Terms: speech analysis, fundamental frequency, periodicity extraction, pathological voice

10.21437/Interspeech.2013-21

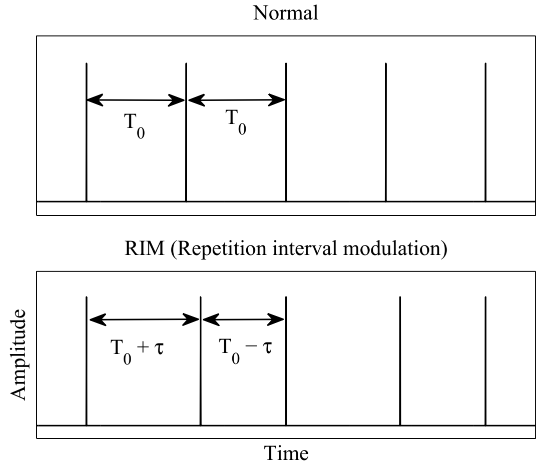

Figure 1: Characteristic structure caused by repetition interval modulation.

1. Introduction

The fundamental frequency (F0) of a voiced sound is defined as the shortest period of the glottal vibrations, and it is one of the indispensable parameters in speech processing such as speech synthesis [1]. Almost all the conventional methods, for example the Cepstrum [2, 3], autocorrelation based method [4], and average magnitude difference function (AMDF) based methods [5], have tried to extract a specific periodic structure represented by a single F0 [6]. More recently, not only new features such as instantaneous frequency [7] and waveform symmetry [8], but also new approaches such as YIN [9], SWIPE [10], neural network based method [11], and combination of these features [12] have been proposed to improve the estimation accuracy. However, some characteristic voices such as the pathological voice [13] and the singing voice [14] often contain a complex excitation behavior represented by multiple periodicity. However, conventional speech processing methods do not deal with multiple periodicity. Being able to analyze and synthesize them would therefore provide a new technique for expanding speech processing [15].

A conventional method [16] has been proposed that deals with multiple periodicities by adding several periodic signals with different F0. On the other hand, the multiple periodicity using a repetition interval modulation (RIM) of the glottal vibrations has been observed for pathological voices [17]. In pathological voices, events such as vocal fold closure seem random, but there are some structures in interval between events even in such voices. For example, in diplophonia, each short interval is coupled with a long interval to form a longer unit. It is a source of subharmonic component. This paper proposes a method for extracting the RIM of glottal vibrations to deal with these types of excitation behavior. The proposed method is based on the TANDEM-STRAIGHT [18] idea, which can estimate the temporally stable power spectrum. The performance of the proposed method is verified by using several artificial signals and pathological voices recorded in a real environment, and the effectiveness of the proposed method is also discussed.

2. Definition of the multiple periodicity

This section discusses the definition of the multiple periodicities caused by RIM. It is defined as the time shift of every other glottal vibration, as shown in Fig. 1. The signal in Fig. 1 represents an example, where three period ($T_0 + \tau$, $T_0 - \tau$, and $2T_0$ as subharmonics) are extracted at the same time.

The proposed method extracts all three periods as candidate intervals and assigns each periodicity score representing salience of repetition. When the waveform shifted to the time of a candidate interval is close to the original waveform, high salience value is assigned. The candidate intervals are rank ordered using associated salience values and the primary one is selected as the best candidate to calculate F0. Details of the definition of the salience are given in the following sections.

3. Proposed method

Figure 2 illustrates the structure of the proposed method. It uses many detectors to calculate the F0 candidates and saliences. Since one detector is designed to extract them in a narrow fre-

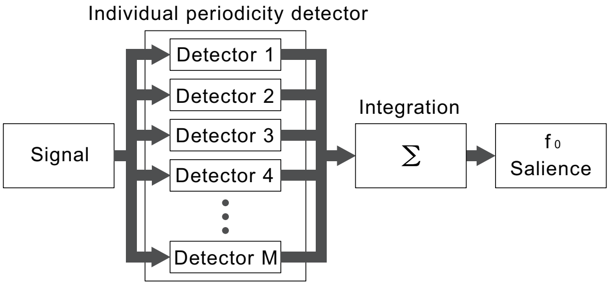

Figure 2: Structure of proposed method.

quency band, many detectors are used to cover a wide range of frequency bands.

3.1. Basic theory

The proposed method is based on the method to estimate a temporally stable power spectrum and spectral envelope [18]. The temporally stable power spectrum $P_T(\omega, t)$ is given by

$$P_T(\omega, t) = \frac{1}{2}\left(P\left(\omega, t - \frac{T_0}{4}\right) + P\left(\omega, t + \frac{T_0}{4}\right)\right), \quad (1)$$

where $P(\omega, t)$ represents the power spectrum of the waveform windowed at time t. T_0 represents the fundamental period ($= 1/f_0$), and ω represents the angular frequency. TANDEM-STRAIGHT uses a pitch synchronous analysis [19] to remove the time varying component and to obtain the temporally stable power spectrum. Since the $P_T(\omega, t)$ estimated by using TANDEM-STRAIGHT is constant without relying on the temporal position for windowing [18], $P_T(\omega)$ without t is used instead of $P_T(\omega, t)$ in the following discussions.

A smoothed power spectrum $P_S(\omega)$ is given by the following equation.

$$P_S(\omega) = \frac{1}{\omega_0}\int_{-\frac{\omega_0}{2}}^{\frac{\omega_0}{2}} P_T(\omega - \lambda)d\lambda, \quad (2)$$

where ω_0 represents the fundamental angular frequency ($=2\pi f_0$). $P_S(\omega)$ is the spectrum that does not contain any periodicity information. On the other hand, $P_T(\omega)$ consists of not only the periodicity information but also the spectral envelope information. Therefore, $P_P(\omega)$, which is the basis of the detector, is given by

$$P_P(\omega) = \frac{P_T(\omega)}{P_S(\omega)} - 1. \quad (3)$$

$P_T(\omega)/P_S(\omega)$ contains a bias that is removed by subtracting 1. $P_P(\omega)$ contains only the periodicity information and equals a sine wave $\cos(2\pi\omega/\omega_0)$ [18]. A unique peak at T_0 is therefore observed by using the inverse Fourier transform of $P_P(\omega)$.

3.2. Detection of F0 and salience

Since F0 in the input voice is unknown before processing, a frequency $f_c(= 1/T_c)$ is used to calculate the F0 candidates and saliences in a given frequency band around f_c. The detector is defined as the inverse Fourier transform of $P_P(\omega; f_c)$. When the target F0 is close to f_c, the detector peaks at the target T_0. Many detectors are therefore used to detect a wide range of F0.

The energy of a voice in a high frequency band is lower than that in a low frequency band, and using all the frequency bands causes a decrease of performance. Before using the inverse Fourier transform, $P_P(\omega; f_c)$ is weighted to reduce the error in the high frequency band. The detector $r_A(\tau; f_c)$ is given by the following equation.

$$\begin{aligned} r_A(\tau; f_c) &= \int_{-\infty}^{\infty} w_B(\omega; f_c)P_P(\omega; f_c)e^{j\omega\tau}d\omega, \quad (4)\\ w_B(\omega; f_c) &= \begin{cases} 1 + \cos\left(\frac{\pi\omega}{N_h\omega_c}\right) & |\omega| \le N_h\omega_c \\ 0 & |\omega| > N_h\omega_c, \end{cases} \end{aligned}$$

where ω_c represents the angular frequency of f_c, and N_h represents the number of harmonics used for weighting. In the proposed method, $r_A(\tau; f_c)$ is converted into the spectral representation $r_A(f; f_c) = r_A(1/\tau; f_c)$.

The frequencies that indicate the peaks in a detector $r_A(f; f_c)$ show the F0 candidates, and the values at the frequencies represent the saliences. For f_c equal to the target F0, the $r_A(f; f_c)$ value at the target F0 is maximum, and the $r_A(f; f_c)$ value decreases in proportion to the difference between f_c and the target F0. The following processes are carried out to maintain the saliences equality without relying on the difference between f_c and the target F0.

In this paper, $r_A(f; f_c)$ is shaped into a raised cosine shape. The bandwidths of a raised cosine are determined so that the overlap of each detector is 50%. The shape is modified to meet the requirement given by the following equation.

$$\begin{aligned} w_L(\nu) &= \begin{cases} 0.5 + 0.5\cos(\pi\nu L) & |L\nu| \le 1 \\ 0 & |L\nu| > 1, \end{cases} \quad (5)\\ \nu(f; f_c) &= \log_2\left(\frac{f}{f_c}\right), \end{aligned}$$

where L represents the number of detectors per one octave and ν represents the logarithmic frequency normalized by f_c. ν and $r_A(\nu; f_c)$ are introduced to normalize the bandwidth of the window function because the length of the window function depends on F0 in TANDEM-STRAIGHT.

The shape of the detector is modified by the following equation.

$$\begin{aligned} r(\nu) &= e^{\alpha\nu}(w_L(\nu) + \beta w_S(\nu))r_A(\nu), \quad (6)\\ w_S(\nu) &= \begin{cases} 0.5 - 0.5\cos(2\pi\nu L) & |L\nu| \le 1 \\ 0 & |L\nu| > 1, \end{cases} \end{aligned}$$

where α and β are the parameters to shape the detector. The salience is normalized to indicate one when the input signal is a pulse train.

All F0 candidates and saliences are obtained by adding all the modified detectors.

$$r_I(\nu) = \sum_{k\in F_c} r(\nu; f_c(k)), \quad (7)$$

where $r(\nu; f_c(k))$ represents the output of the kth detector. The indexes that have peaks are F0 candidates, and the values for the F0 candidates represent saliences. Since $r_I(\nu)$ has many peaks, a threshold is used to remove the unwanted peaks that were fortuitously caused.

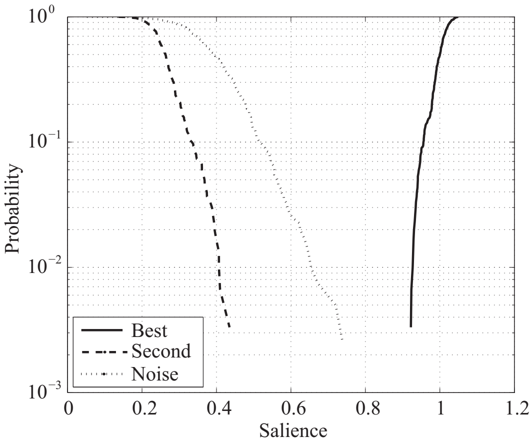

Figure 3: Distribution of periodicity score's peak value. Solid line: primary peak for periodic signal, dashed line: second peak for periodic signal, and dotted line: primary peak for white noise.

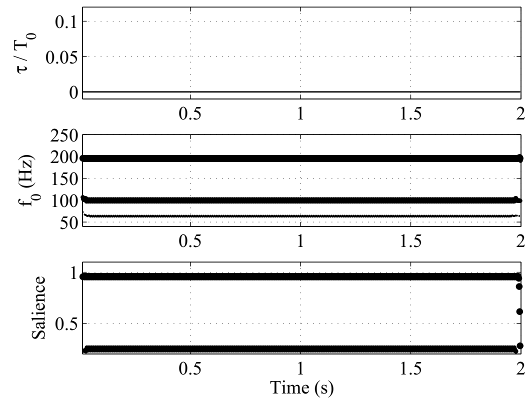

Figure 4: Analysis results of proposed method (F0: 200 Hz, maximum ratio: 0%)

3.3. Determination of parameters

The proposed method has several parameters that affect the extraction performance. In this article, the parameters are determined by the results of a preliminary experiment. These parameters are determined to meet the noise and RIM tolerance. White noise and periodic signals with many different F0s including various RIM patterns are used for the experiment. The number of detector M is set to 15 (3 ch/oct. from 32 to 812.7 Hz). As a result, these values are determined; the number of detectors per octave L is three, the window function is a Blackman window with the length of $2.5T_c$, the number of harmonics N_h is eight, α is 1.0132, and β is 0.0364.

The threshold to remove the unwanted peaks is determined by conducting another experiment. Figure 3 illustrates the cumulative distribution. The test signals are white noise and the periodic signals with several F0s randomly given from 30 to 1000 Hz. The horizontal axis represents the salience used for the threshold, and the vertical axis represents the probability that the peak exceeds the threshold (solid line) or falls below it (dashed and dotted line). The solid line represents the cumulative distribution of the primary peak for the periodic signal, and the dashed one represents that of the second peak for the periodic signal. The dotted line represents that of the primary peak in the white noise.

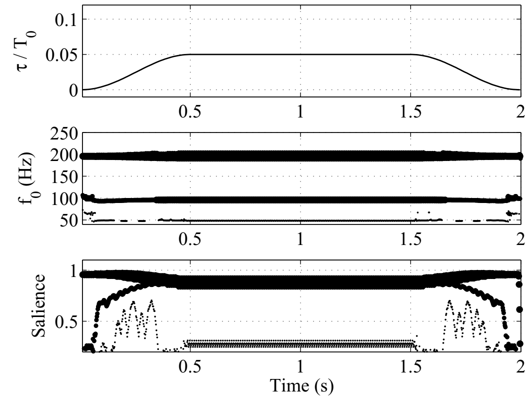

Figure 5: Analysis results of proposed method (F0: 200 Hz, maximum ratio: 5%)

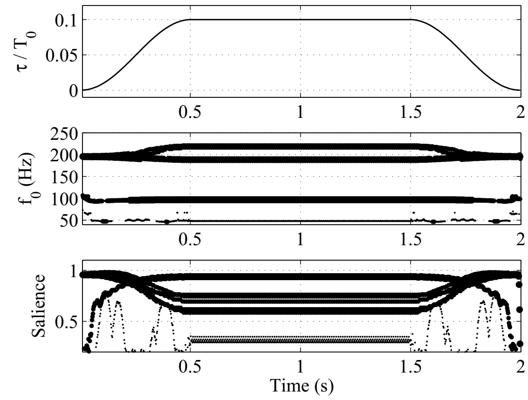

Figure 6: Analysis results of proposed method (F0: 200 Hz, maximum ratio: 10%)

Figure 3 shows that the salience of the second peak for the periodic signal is lower than that of the primary peak for the white noise, which suggests that the proposed method can remove the unwanted peaks by appropriately setting the threshold. When the threshold is set to 0.6, the probability that a wrong F0 is fortuitously detected is 2.7%

4. Evaluation

Two experiments were carried out to demonstrate the effectiveness of the proposed method. In the first experiment, several artificial signals were used, while two pathological voices recorded in a real environment were used in the second experiment. The effectiveness of the proposed method was verified by both experiments.

4.1. Experiment by artificial signals

The basic F0 is set to 200 Hz, and three modulation patterns are used. Modulation patterns are based on the maximum ratio of τ/T_0, and 0% (without modulation), 5%, and 10% are used.

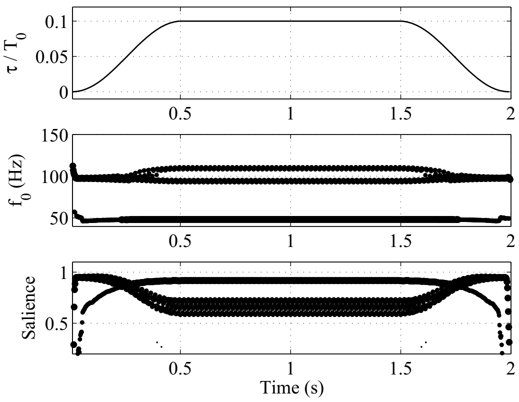

Figure 7: Analysis results of proposed method (F0: 100 Hz, maximum ratio: 10%)

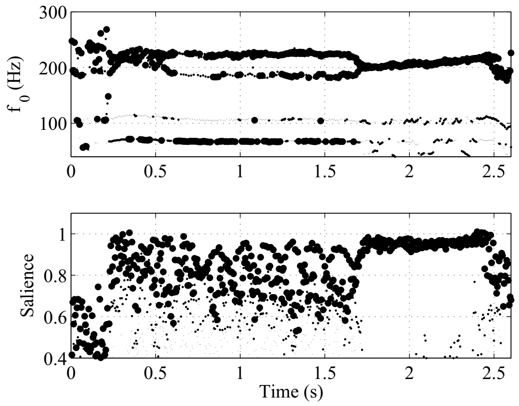

Figure 8: Extracted periodic components by using proposed method. The input voice was a vowel /e/ spoken by a female patient with polypoid vocal folds.

Another condition (F0: 100 Hz, maximum ratio: 10%) is used to verify the dependence in F0.

Figs. 4, 5, 6 and 7 illustrate the analysis results. The middle represents the F0 candidates, the bottom represents saliences, and the marker size of F0 represents the size of salience. In Fig. 4, only the target F0 (200 Hz) is extracted with high salience close to 1.0. In Figs. 5 and 6, the three F0s (Two F0s around 200 Hz and one at 100 Hz as the subharmonics) are extracted, and most of the saliences of these F0s are more than 0.6, which is the threshold necessary to remove the unwanted peak. In Fig. 7, same tendency is observed even if the F0 is different. These suggest that the proposed method can extract RIM at a high level of accuracy.

4.2. Evaluation by the pathological voices

An evaluation using pathological voices recorded in the real environment was carried out. The voices used were the pathological voices attached to a Japanese book [13] and selected by a listening test of sound quality. The sampling of these voices was at 44.1 kHz and 16 bits. This article shows the results of typical two patterns to discuss the effectiveness of the proposed method.

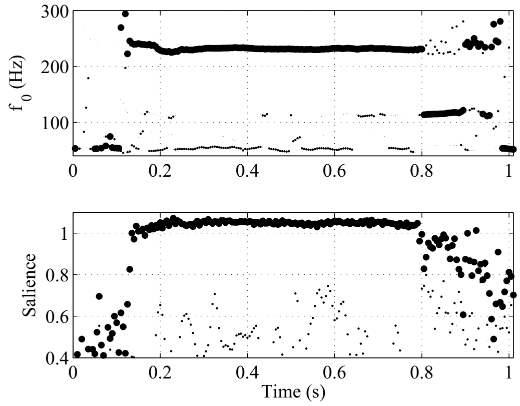

Figure 9: Extracted periodic components using proposed method. The input voice was a vowel /i/ spoken by a female patient with scarred vocal folds.

Figure 8 illustrates the results for the voices pronouncing the vowel /e/ spoken by a female patient with polypoid vocal folds. The marker size represents the size of the salience. The F0 candidates around 200 Hz and the subharmonic of 100 Hz are observed in the timeframe from 0.5 to 1.5 sec, and saliences were indicated around 0.8. Figure 9 illustrates the results for the voices pronouncing the vowel /i/ spoken by a female patient with scarred vocal folds. This figure also indicates the three F0s in the timeframe from 0.8 to 0.9 sec.

4.3. Discussion

These results seem to show that the pathological voices used in the experiment contain multiple periodicities caused by RIM. The saliences in Figs. 8 and 9 spread at a wide range from 0.4 to 1.0. However, since most of the F0 candidates that should be observed were extracted with more than 0.5 saliences, we can manually extract the RIM without needing to carefully watch the waveform, which suggests that the proposed method is useful as a diagnostic tool to extract such pathological voices in the medical field. This paper does not deal with the automation of the extraction, and this is one of our more important future works.

5. Conclusions

We proposed a method for extracting the multiple periodicities caused by RIM from a voiced sound. The proposed method uses the TANDEM-STRAIGHT idea and extracts several F0 candidates and saliences. In this article, the voice experiments including RIM were carried out to verify the effectiveness of the proposed method. The analysis results from the artificial signal indicated that the proposed method can extract typical F0s caused by RIM, and the analysis results of pathological voices show results similar to those from the artificial signal. The results suggested that the proposed method can observe the RIM from the three F0s caused by RIM.

6. Acknowledgment

The authors would like to thank Ms. Y. Wada for valuable discussions. This work was supported by JSPS KAKENHI Grant Numbers 23700221, 23500233, 24300073, and 24650085.

7. References

[1] H. Dudley, “Remaking speech,” *J. Acoust. Soc. Am.*, vol. 11, no. 2, pp. 169–177, 1939.

[2] A.M Noll, “Short-time spectrum and “cepstrum” techniques for vocal pitch detection,” *J. Acoust. Soc. Am.*, vol. 36, no. 2, pp. 269–302, 1964.

[3] A.M Noll, “Cepstrum pitch determination,” *J. Acoust. Soc. Am.*, vol. 41, no. 2, pp. 269–309, 1967.

[4] L.R. Rabiner, “On the use of autocorrelation analysis for pitch detection,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 1, pp. 24–33, 1977.

[5] M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, and H.J. Manley, “Average magnitude difference function pitch extractor,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. ASSP-22, no. 5, pp. 353–362, 1974.

[6] W. Hess, *Pitch determination of speech signals*, Springer-Verlag, 1983.

[7] T. Abe, T. Kobayashi, and S. Imai, “Harmonics tracking and pitch extraction based on instantaneous frequency,” in *Proc. ICASSP95*, pp. 756–759, 1995.

[8] H. Kawahara, M. Morise, R. Nisimura, and T. Irino “Deviation measure of waveform symmetry and its application to high-speed and temporally-fine F0 extraction for vocal sound texture manipulation,” in *Proc. INTERSPEECH2012*, 4-page, 2012.

[9] A.de. Cheveigne, and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[10] A. Camacho, and J.G. Harris, “A sawtooth waveform inspired pitch estimator for speech and music” *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[11] E. Barnard, R.A. Cole, M. P. Vea, and F.A. Alleva, “Pitch detection with a neural-net classifier,” *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 298–307, 1991.

[12] H. Kawahara, A.de. Cheveigne, H. Banno, T. Takahashi, and T. Irino, “Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT,” in *Proc. INTERSPEECH2005*, pp. 537–540, 2005.

[13] The Japan Society of Logopedics and Phoniatrics, Interuna Publisher, Inc. ISBN:978-4-900637-21-4, 2005 (in Japanese).

[14] O Fujimura, K. Honda, H. Kawahara, Y. Konparu, M. Morise, and J.C. Williams, “Noh voice quality,” *Logopedics Phoniatrics Vecology*, vol. 34, pp. 157–170, 2009.

[15] H. Kawahara, M. Morise, and T. Irino, “Analysis and synthesis of strong vocal expressions: Extension and application of audio texture features to singing voice,” in *Proc. ICASSP2012*, pp. 5389–5392, 2012.

[16] A.de. Cheveigne and H. Kawahara, “Multiple period estimation and pitch perception model,” *Speech Communication*, vol. 27, pp. 175–185, 1999.

[17] H. Itagaki, M. Morise, R. Nisimura, T. Irino, and H. Kawahara, “A bottom-up procedure to extract periodicity structure of voiced sounds and its application to represent and restoration of pathological voices,” in *Proc. MAVEBA09*, pp. 115–118, 2009.

[18] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, “TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation,” in *Proc. ICASSP2008*, pp. 3933–3936, 2008.

[19] M.V. Mathews, J.E. Miller, and E.E. David, “Pitch synchronous analysis of voiced sounds,” *J. Acoust. Soc. Am.*, vol. 33, pp. 179–185, 1961.