CHARACTERISATION OF TREMOR IN NORMOPHONIC VOICES

Rubén Fraile, Nicolás Sáenz-Lechón, Víctor J. Osma-Ruiz, Juana M. Gutiérrez-Arriola

Signal Theory & Communications Department - Universidad Politécnica de Madrid

ABSTRACT

Vocal tremor is a low frequency instability of the voice that causes modulation of its amplitude and fundamental frequency. Among these two, frequency modulation is more relevant for perception and it has been shown to be present both in normophonic and dysphonic voices and to happen in similar frequency bands for both voice types. This paper presents a characterisation of the frequency modulating signal estimated for normophonic voices in terms of both its spectral characteristics and its statistical distribution. By using the discrete Fourier transform for data non-uniformly spaced in time domain, it is shown that the modulating signal may be either low-pass or band-pass (i.e. oscillating), though the low-pass case dominates in the analysed data. As for the values of the modulating signal, their distribution is shown to fairly fit a Gaussian distribution with a standard deviation that significantly depends on the average fundamental frequency.

Index Terms— Acoustic signal analysis, Biomedical acoustics, Frequency modulation, Speech analysis

1. INTRODUCTION

Vocal tremor may be defined as a low frequency (1 to 15 Hz) modulation of either the amplitude or the fundamental frequency of the voice signal [1]. While both kinds of modulation (amplitude and frequency) coexist, only frequency modulation seems to be relevant for perception [2, 3]. Such modulation is present in voices from both healhty (normophonic) and dysphonic speakers and it has been measured to happen in similar (if not the same) frequency bands for both situations [2, 4, 5]. In fact, as far as tremor is concerned, the difference between normophonic and dysphonic voices has been reported to be more in the modulation extent than in the modulation rate [2, 4].

Regarding the modulation rate or, better, the spectral characteristics of the modulating signal, tremor is frequently assumed to be oscillatory [6, 7], which corresponds to a band-pass spectrum for the modulating signal. Yet, some results indicate that tremor may not be oscillatory [3,8] and, as a consequence, the modulating signal may be low-pass. As for the statistical distribution of the values of the modulating signal, to the best of authors' knowledge, it has not been analysed to present.

This work has been carried out in the framework of project grant TEC2012-38630-C04-01, financed by the Spanish Government.

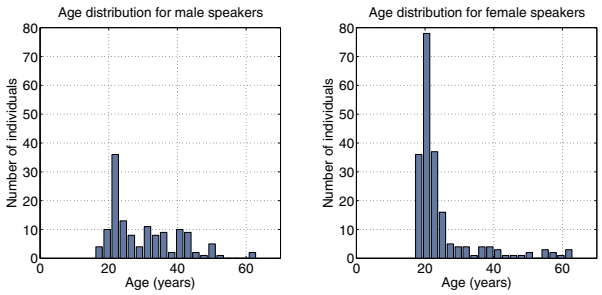

Fig. 1. Age distribution of speakers by sex.

This paper aims at characterising the frequency modulating signal associated to tremor in terms of both its spectral characteristics and the statistical distribution of its values. Following some previous approaches, the acoustic analysis of vocal tremor is performed by studying the sequence of pitch periods or pitch process (e.g. [5, 9]). The analysis procedure is similar to the one in [5] with some variations. Details are included in Section 3. One of the most relevant of such variations is the use of the discrete Fourier transform for data non-uniformly spaced in time domain; such use is justified by the nature of the pitch process [9].

2. MATERIALS

341 recordings were chosen among those corresponding to the phonation of vowel */a/* at normal pitch in the Saarbruecken voice database [10]. The choice of recordings corresponding to the phonation of sustained vowels is justified by the fact that tremor is better perceived in sustained phonations than in running text [11]. For all recordings, the sampling frequency was equal to 50 kHz and the number of quantization bits was 16.

Recordings were selected according to the criteria of corresponding to healthy individuals and having a duration of at least 1 second, hence allowing a resolution as low as 1 Hz in spectral domain. Each selected recording corresponded to one different speaker. Among the 341 speakers, there were 135 males and 206 females. The distribution of ages for each sex is depicted in Figure 1.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

3. METHODS

All voice recordings were cut before processing in order to have the same duration for all of them (1 second). The use of signal durations for vocal tremor analysis in the range from 1 to 2 seconds has been previously reported in the literature [4, 5, 8]. Voice onsets and offsets were avoided in all cases.

3.1. Estimation of the average pitch period

Average pitch period estimation of the 1-second signals was based on the autocorrelation function of each voice signal. Firstly, the 16-bit signal was converted to 2 bits by assigning the value -1 to negative samples of the original recording and the value +1 to positive samples. This procedure was used in [5] and it helps to emphasize the most prominent (i.e. wide) peaks of the signal waveform [12, p.154]. A biased estimate of the autocorrelation function was preferred in order to avoid undesirable peaks at the tails of the function:

$$\hat{r}_{xx}[m] = \frac{1}{N}\sum_{n=m}^{N-1} x_{\mathrm{b}}[n]\, x_{\mathrm{b}}[n-m] \tag{1}$$

where $x_{\mathrm{b}}[n]$ is the binary version of the discrete-time voice signal $x[n]$, f_s is the sampling rate, N is the signal duration in number of samples and $\frac{m}{f_s}$ is the time lag.

Once the autocorrelation estimate was available, the positions of its local maxima m_i $i \in \{1, 2 \dots M\}$ were ordered by their corresponding autocorrelation values $\hat{r}_{xx}[m_i]$ in descending order. The first value of m_i was then analysed so as to check if there existed other maxima in the intervals $2m_i \pm 0.1m_i$ and $3m_i \pm 0.1m_i$. In this event, $\hat{T}_0 = \frac{m_i}{f_s}$ was taken to be the average pitch period of the signal. If the condition was not met, then the value of m_i corresponding to the next local maximum was checked.

In case a suitable estimate in the range $\frac{1}{\hat{T}_0} \in [50, 700]$ could not be found, then $x_{\mathrm{b}}[n]$ was split into three non-overlapping segments with lengths equal to $\lfloor \frac{N}{3} \rfloor$, the procedure was repeated independently for each segment and the median estimate of $\hat{T}_0$ was kept as the average value for the whole signal.

3.2. Location of the individual pitch periods

The estimated average pitch period was subsequently used as a reference for the location of the individual pitch periods within the voice signal $x[n]$. Previously, the signal was upsampled by repetition of its samples twice, so as to get a sampling rate equal to $f'_s = 150$ kHz. This procedure had the purpose of improving the time resolution of the pitch length estimation results [5]. After that, a linear-phase band-pass filter was applied with cut-off frequencies equal to 300 Hz and 3 kHz. The rationale for low-pass filtering is given in [5]. However, a higher cut-off frequency was preferred in order to smooth less the most prominent peaks of the signal. The low frequencies were also attenuated in order to remove additive artefacts slower than the voice fundamental frequency that could pose difficulties in the identification of corresponding peaks in consecutive periods. The resulting filter had the same pass band as the analogue telephone channel [13]. Indeed, it has been shown that the signal distortion caused by the telephone channel has little impact on the capacity to detect tremor [14].

The highest maximum in the 16-bit band-pass signal $x'[n]$ was used as a reference point for starting the iterative identification of pitch periods. Given the position m_M of that maximum, a signal frame around it was selected having a length approximately equal to one fifth of the average pitch period N_{P}:

$$x_{m_M}[n] = x'[n - m_M],\ |n| \le N_{\mathrm{f}},\ N_{\mathrm{f}} = \lfloor 0.1 N_{\mathrm{P}} \rfloor \tag{2}$$

where $N_{\mathrm{P}} = \lfloor \hat{T}_0 f'_s \rfloor$. Afterwards, a corresponding frame was selected one period to the left of m_M and the cross-correlation between both frames estimated:

$$\hat{r}'_{xx}[q] = \frac{\sum_{n=\max\{-N_{\mathrm{f}}, q-N_{\mathrm{f}}\}}^{\min\{N_{\mathrm{f}}, -q+N_{\mathrm{f}}\}} x_{m_M}[n]\, x_{m_M - N_{\mathrm{f}}}[n-q]}{2N_{\mathrm{f}} + 1} \tag{3}$$

If q_M was the position of the highest peak of $\hat{r}'_{xx}[q]$ for $-2N_{\mathrm{f}} \le q \le 2N_{\mathrm{f}}$, then the position of the peak corresponding to m_M in the previous pitch period was estimated as $m_M - N_{\mathrm{P}} - q_M$ and its corresponding period duration, as $\tau_k = \frac{(N_{\mathrm{P}} + q_M)}{f'_s}$. The full sequence of τ_k values for the signal was obtained by iterative repetition of this procedure first to the left of m_M and afterwards to the right. Note that the resolution of this algorithm in the estimation of τ_k is equal to $1/f'_s$. With respect to [5], this procedure for the identification of individual pitch periods was found to be more robust, since the correlation considers not only the individual values of the signal at its local maxima, as peak picking algorithms do, but also a certain environment around them.

3.3. Estimation of the spectrum of the pitch process

The pitch period estimates τ_j were subsequently inverted to obtain a series of fundamental frequency estimates $f_k = \frac{1}{\tau_k}$. The sequence of such estimates can be viewed as a series of samples of a continuous-time signal $f(t) = \frac{1}{\hat{T}_0} + \theta(t)$ taken at non-equispaced time instants $t_k = \sum_{l=1}^{k} \tau_l$. The spectrum of the modulating signal $\theta(t)$ was thus estimated by making use of the non-equispaced discrete-time Fourier transform (derived from [15]):

$$\Theta(\omega) = \sum_{k=1}^{K} \left(f_k - \frac{1}{\hat{T}_0} \right) \mathrm{e}^{-j\omega t_k} \tag{4}$$

where K is the number of pitch periods in $x[n]$.

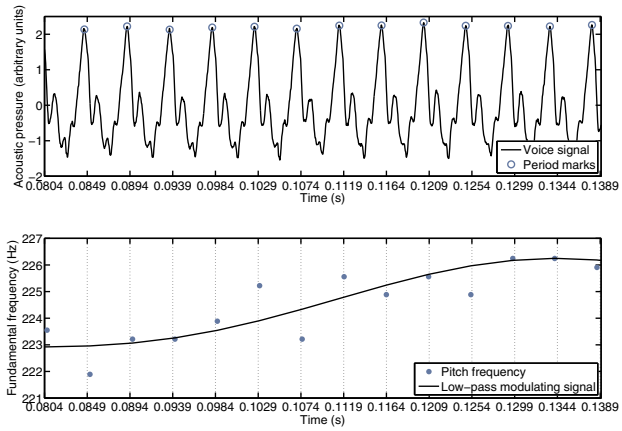

Fig. 2. Estimated pitch periods (up) and corresponding fundamental frequencies and low-pass modulating signal (down). It can be noticed that pitch periods at the left of the graph are longer than at the right.

A non-equispaced or non-uniform discrete Fourier transform (NDFT) was obtained by sampling $\Theta(\omega)$ at regularly spaced angular frequencies $\omega_a = \frac{2\pi a}{K\overline{\tau}}$, where $a \in \{0, 1 \ldots K-1\}$ and $\overline{\tau} = \frac{\tau_K}{K-1}$:

$$\Theta(\omega_a) = \sum_{k=1}^{K} \left(f_k - \frac{1}{\hat{T}_0} \right) \mathrm{e}^{-j2\pi \frac{a}{K} \frac{t_k}{\overline{\tau}}} \tag{5}$$

Last, the spectral components were estimated by squaring the modulus of $\Theta(\omega_a)$ and the final spectrum estimate was obtained by smoothing $|\Theta(\omega_a)|^2$ using a Hamming window with length equivalent to 20 Hz.

3.4. Identification of the relevant spectral components

Once an estimate of the spectral components $\Theta(\omega_a)$ of the zero-average frequency modulating signal $\theta(t)$ was available, the relevance of each of them was assessed by using the non-parametric method described in [5]. Such method involved random reordering of the sequence f_k, spectrum estimation for the reordered sequence, and discarding all the components ω_a whose corresponding estimated power density $|\Theta(\omega_a)|^2$ after smoothing was lower for the original sequence than for the shuffled one. By repeating the process a sufficient number of times (250 in this experiment), only the relevant components of the pitch process spectrum were kept. Among them, those belonging to the interval between 1 and 25 Hz ($1 \leq \frac{\omega_a}{2\pi} \leq 25$) were selected as being associated to vocal tremor.

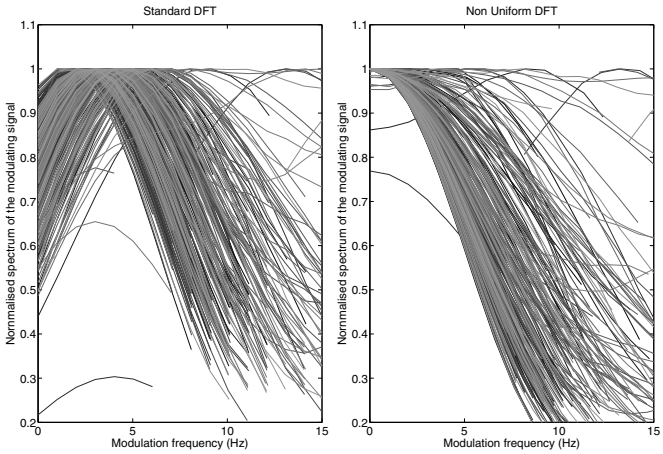

Fig. 3. Relevant components of the spectrum of the modulating signal $\theta(t)$, estimated using the NDFT (right) and the standard DFT (left).

4. RESULTS

For illustration purposes, Figure 2 (top) shows the result of applying the pitch identification procedure to a sample voice signal. The corresponding pitch frequencies and the frequency modulating signal obtained by computation of the inverse NDFT of the spectral components associated to vocal tremor (*low-pass modulating signal*) are depicted at the bottom graph of the figure.

4.1. Spectral characteristics of the modulating signal

Figure 3 shows the estimated spectrum for the modulating signal $\theta(t)$ corresponding to each of the 341 voice recordings available. The spectrum estimates obtained using the NDFT are plot on the right graph while spectrum estimates obtained using the standard DFT have been plot on the left for the sake of comparison. It can be noticed that the assumption of uniform sampling of $\theta(t)$, implicit in the estimation using the standard DFT, produces a spectrum estimate shifted to the right of the frequency axis. This implies that the resultant spectrum for most of the cases is band-pass, while the spectrum estimates obtained via the NDFT are low-pass for the majority of signals.

The difference between both sets of results may be further appreciated in Figure 4, where the average modulation frequency calculated for each voice recording has been plotted as a function of the peak modulation frequency. The peak modulation frequencies were obtained by simply detecting the maxima in Figure 3, while the average modulation frequencies were calculated as:

$$\overline{f}_{\mathrm{mod}} = \frac{\sum_{a \in A_r} \frac{\omega_a}{2\pi} |\Theta(\omega_a)|^2}{\sum_{a \in A_r} |\Theta(\omega_a)|^2} \tag{6}$$

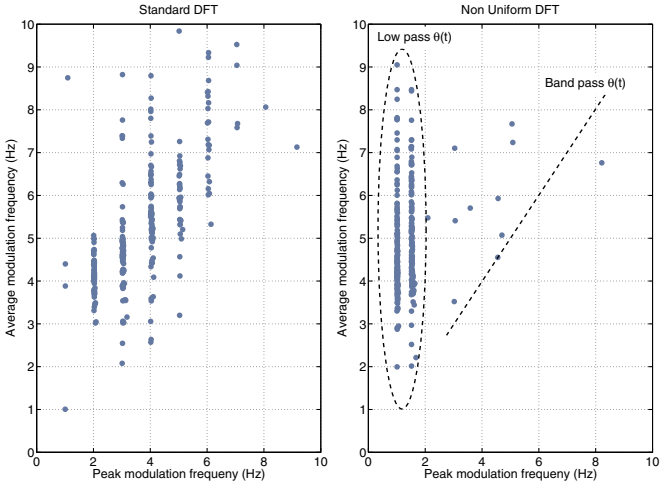

Fig. 4. Scatter plot of average modulation frequency vs peak modulation frequency, estimated using the NDFT (right) and the standard DFT (left).

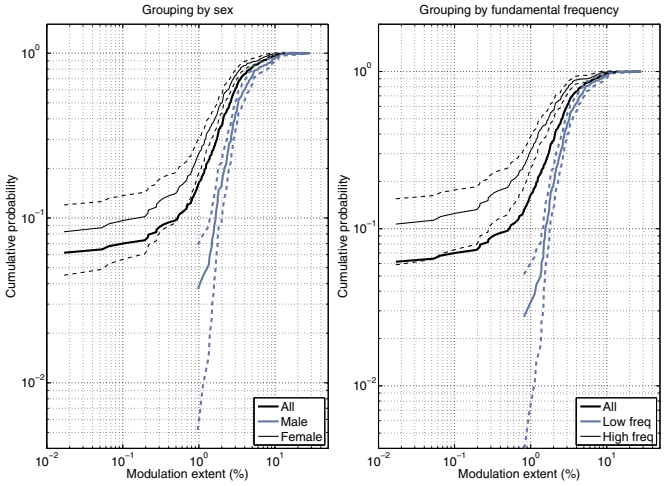

Fig. 5. Cumulative distribution functions (CDF) of measured modulation extents. In the left graph, signals are separated into two classes according to speakers' sex (male or female). The thick blak line indicates the overall CDF. Dashed lines indicate the 95% confidence interval for each CDF. In the right graph, signals are separated according to their average fundamental frequency (high or low). The threshold between both groups was calculated according to Otsu's criterion [16].

where A_r is the subset of $\{0, 1 \ldots K-1\}$ corresponding to relevant frequencies between 1 and 25 Hz.

By examining Figure 4 (right), it can be noticed that the modulating signals can be classified as belonging to two distinct groups: those having the highest spectral peak at low frequencies, independently of their average frequency, and those for which their average frequency approximately is a growing function of the frequency of the highest spectral peak. The first group corresponds to low-pass modulating signals while the second one corresponds to band-pass modulating signals. Such difference does not exist when the spectrum is estimated using the standard DFT (Figure 4 (left)).

4.2. Distribution of modulation extents

After identification of the relevant spectral components between 1 and 25 Hz, the low-pass modulating signal $\theta_{\mathrm{LP}}(t)$ was estimated by calculating the inverse NDFT of $\Theta(\omega_a)$ for $a \in A_r$. From $\theta_{\mathrm{LP}}(t)$, the relative modulation extent for each voice signal was measured as:

$$\Delta f_{\mathrm{mod}} = \left(\max\left\{\theta_{\mathrm{LP}}(t)\right\} - \min\left\{\theta_{\mathrm{LP}}(t)\right\}\right) \cdot \hat{T}_0 \quad (7)$$

Figure 5 shows the experimental cumulative distributions of $\Theta(\omega_a)$ when signals are grouped by speakers' sex (left) and by fundamental frequency (right). Overall, the modulation extent is less than 11% for 99% of cases. However, the distributions of modulation extents for male and female speakers are significantly different ($p < 0.001$ for a Wilcoxon test of the mean [17]). Such significant difference also happens when discriminating voices by fundamental frequency (Figure 5 (right)).

4.3. Distribution of fundamental frequencies

Figure 6 (left) shows the probability plot of the normalised values of $\theta_{\mathrm{LP}}(t)$ for all voice signals. These values have been obtained by normalising each modulating signal independently:

$$\overline{\theta}_{\mathrm{LP}}(t) = \frac{\theta_{\mathrm{LP}}(t) - \mathrm{E}\left\{\theta_{\mathrm{LP}}(t)\right\}}{\sigma_\theta} \quad (8)$$

being σ_θ the standard deviation of $\theta_{\mathrm{LP}}(t)$. The probability plot indicates a reasonably good fit between the normal (i.e. Gaussian) distribution and the normalised distribution of $\theta_{\mathrm{LP}}(t)$. The experimental CDF of $\sigma_\theta \cdot \hat{T}_0$ is plotted in Figure 6 (right). Values below 2.4% have been measured for 99% of cases, although high-pitched voices have values significantly lower than low-pitched voices.

5. CONCLUSIONS

The spectral and statistical characteristics of vocal tremor in normophonic voices have been analysed. Specifically, the frequency modulation present in 341 phonations of vowel */a/* at normal pitch has been studied.

For the processed dataset, the average modulation frequencies range from 1 to 10 Hz (Figure 4). When evaluating the shape of the spectrum of the modulating signals, it has been shown that using the standard DFT may lead to erroneous conclusions since the nature of the pitch process is that

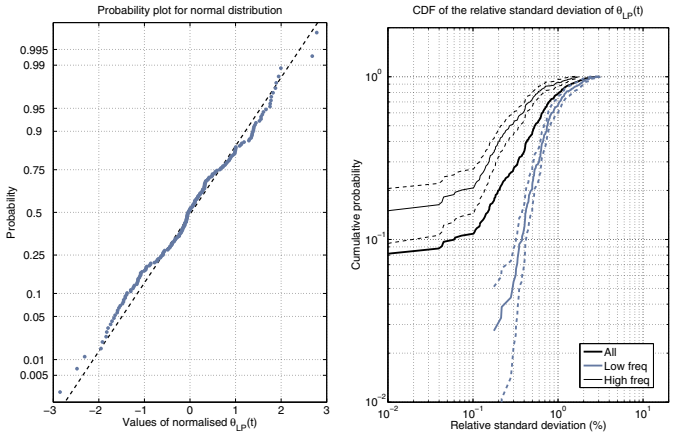

Fig. 6. Probability plot of the normalised aggregate distribution of $\theta_{\mathrm{LP}}(t)$ (left). CDF of measured relative standard deviations of $\theta_{\mathrm{LP}}(t)$ (right).

of a non-uniformly sampled discrete signal. The use of the NDFT for characterising the modulating signals leads to the conclusion that these mostly are low-pass though band-pass modulations were also found.

As for the modulation extents, 99% of the analysed voices have values below 11% relative to their average fundamental frequency (Figure 5). Yet, a significant difference exists between male and female speakers. Such a difference is likely to be related to the average fundamental frequency: low-pitched voices tend to experience modulations with higher relative extents than high-pitched voices. The relation between fundamental frequency and tremor had also been detected by Dromey et al. [8], but they focused on variations for the same speaker and herein inter-speaker differences have been detected (speakers with lower pitch frequency tend to have higher tremor extents than speakers with higher pitch frequency).

Last, it has been shown that the distribution of fundamental frequencies associated to pitch processes can be modelled by normal (Gaussian) random variables (Figure 6) with normalised standard deviations that in 99% of cases have values below 2.4% of the fundamental frequency. For high-pitched voices ($\frac{1}{T_0} > 174$ Hz), this threshold is even lower (1.6%).

REFERENCES

[1] I.R. Titze, "Workshop on acoustic voice analysis: Summary statement," National Center for Voice and Speech, 1995.

[2] S. Anand, R. Shrivastav, J.M. Wingate, and N.N. Chheda, "An acoustic-perceptual study of vocal tremor," *Journal of Voice*, vol. 26, no. 6, pp. 811–817, 2012.

[3] J. Kreiman, B. Gabelman, and B.R. Gerratt, "Perception of vocal tremor," *Journal of Speech, Language, and Hearing Research*, vol. 46, no. 1, pp. 203–214, 2003.

[4] J. Jiang, E. Lin, and D.G. Hanson, "Acoustic and airflow spectral analysis of voice tremor," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 1, pp. 191–204, 2000.

[5] J. Schoentgen, "Modulation frequency and modulation level owing to vocal microtremor," *Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 690–700, 2002.

[6] G. Deuschl, P. Bain, and M. Brin, "Consensus statement of the Movement Disorder Society on tremor," *Movement Disorders*, vol. 13, no. S3, pp. 2–23, 1998.

[7] A.G. Shaikh, K. Miura, L.M. Optican, S. Ramat, R.M. Tripp, and D.S. Zee, "Hypothetical membrane mechanisms in essential tremor," *Journal of Translational medicine*, vol. 6, no. 1, pp. 68–78, 2008.

[8] C. Dromey, P. Warrick, and J. Irish, "The influence of pitch and loudness changes on the acoustics of vocal tremor," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 5, pp. 879–890, 2002.

[9] I. Gath and E. Yair, "Comparative evaluation of several pitch process models in the detection of vocal tremor," *IEEE Transactions on Biomedical Engineering*, , no. 7, pp. 532–538, 1987.

[10] M. Pützer and W.J. Barry, "Saarbruecken voice database," Online: *www.stimmdatenbank.coli.uni-saarland.de*, Feb. 2015, version 2.0.

[11] A. Lederle, J. Barkmeier-Kraemer, and E. Finnegan, "Perception of vocal tremor during sustained phonation compared with sentence context," *Journal of Voice*, vol. 26, no. 5, pp. 668.e1–e9, 2012.

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

[13] ITU-T, "Transmission characteristics of national networks," Series G: Transmission Systems and Media, Digital Systems and Networks Rec. G.120 (12/98), Dec 1998.

[14] D.M. Johnson, E.R. Hapner, A.M. Klein, M. Pethan, and M.M. Johns, "Validation of a telephone screening tool for spasmodic dysphonia and vocal fold tremor," *Journal of Voice*, vol. 28, no. 6, pp. 711–715, 2014.

[15] D. Potts, G. Steidl, and M. Tasche, "Fast fourier transforms for nonequispaced data: A tutorial," in *Modern sampling theory*, pp. 247–270. Springer, 2001.

[16] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[17] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.