Nine voices, one artist: Linguistic and acoustic analysis.

Talal Bin Amin, Pina Marziliano, James Sneed German

Nine Voices, One Artist: Linguistic and Acoustic Analysis

Talal Bin Amin, Pina Marziliano
School of Electrical and Electronic Engineering,
Nanyang Technological University,
Singapore
talal1@e.ntu.edu.sg, epina@ntu.edu.sg

James Sneed German
School of Humanities and Social Sciences,
Nanyang Technological University,
Singapore
jsgerman@ntu.edu.sg

***Abstract*—Voice impersonators possess a flexible voice and thus can change their voice identity. They are able to imitate various people and characters which differ in age, gender, accent and voice quality. State of the art electronic voice conversion systems are not able to successfully mimic their human counterparts as they lack naturalness. To understand why human impersonators are successful and what parameters they rely on to change their voice, we analyze nine voices produced by a professional voice impersonator. We compute different acoustical measures and discuss their linguistic implications. The acoustical measures include pitch, speech rate and formant frequencies. Our results show that differences in the voice identity features such as age and gender are reflected in the acoustic parameters of the impersonations. The analysis is distinguished from previous studies on impersonators in giving full consideration to voice identity features.**

***Index Terms*—impersonator, language independent, prosodic, voice identity, voice-over artist**

I. Introduction

Voice impersonation involves changing one's voice to sound like another person. It is mostly done for entertainment purposes, e.g. caricaturization and in media related fields. However, the study of voice impersonators is also important in other fields of research including forensics [1], speaker recognition [2] and voice conversion [3]. From the point of view of forensics, for example, one can disguise voice to alter identity and get away with a crime. In that respect it is important to develop measures and techniques that will allow law enforcing authorities to identify a criminal even from the disguised voice itself. Similar issues apply to security systems based on speaker recognition, which are also vulnerable to attacks by voice impersonators. The study of voice impersonators is therefore useful to answer what features they rely on to disguise their voice and how to normalize them. Also, in voice conversion, the goal is to convert the voice of a given source speaker to that of a target speaker. However, current algorithms often lack naturalness as well as individuality [4], where individuality refers to the (lack of) similarity of the converted voice to the target speaker. This results in an unnatural third voice in between that of the source and the target speaker. Therefore, it is important to study voice impersonators to find out how they are able to convert their voices while maintaining both naturalness and individuality simultaneously. This will be useful for further improvement of the current voice conversion algorithms.

Studies on voice impersonators are very limited [5], [6], [7], [8] and different observations were reported with different data sets. For example, in [5] 30 second excerpts of uninterrupted Swedish sentences were analyzed whereas in [7] only two short Japanese sentences were used. In [6], different sentences were used while impersonating different persons and only one word was common to all the different sentences and therefore useful for comparisons. Additionally, the different sentences used in [6] were not emotionally neutral as they were designed to be humorous. In [5] it was concluded that the voice impersonator found it difficult to accurately modify vocal tract characteristics towards the target speaker, whereas in [7] the impersonator was able to modify both the prosodic and vocal tract characteristics towards the target speaker. Disagreement in these results could be attributed to the fact that the impersonators used in these studies had different skill sets and also had different targets to imitate in different languages. In all the previous studies including [5], [6], [7], [8] the goal of the impersonator was to imitate particular persons or target speakers. This is in contrast to our data set where the impersonator tried to imitate certain characters, for example, a "scratchy" old female. This allows the impersonators to fully express their flexibility by impersonating voices they feel comfortable with or have a better control over.

In our study, we try to gain a better understanding as to how the voice artist was able to change the different parameters of voice identity. We have performed an acoustical analysis of the various impersonations. The parameters we have measured and analysed include fundamental frequency (pitch), speaking rate (speech rate) and vowel formant frequencies (F1, F2 and F3). Pitch and speaking rate are prosodic features of the speech signal that may be both speaker-specific and relevant for conveying meaning. Formant frequencies are the primary acoustic correlates of differences between vowels and are primarily determined by shape characteristics of the vocal tract. We also show how these parameters might be affected by language or dialect specific features such as those associated with regional accent. The ultimate goal is to identify the parameter(s) which will be useful for speech synthesis algorithms to produce a variety of different voices from a single source voice. The

TABLE I
ABBREVIATIONS OF THE NINE DIFFERENT VOICES FROM THE VOICE ARTIST.

HPF	High Pitched Female
NF	Nasal Female
OF	Old Female
OM	Old Male
N	Natural
SOF	Scratchy Old Female
YF	Young Female
YG	Young Girl
YM	Young Male

terms voice impersonator and voice-over artist will be used interchangeably from now.

The rest of this paper is organized as follows: Section II describes the speech data set provided by the voice over artist; the analysis and results are presented in Section III; and finally Section IV concludes the paper.

II. SPEECH DATA

The speech data was collected from a professional voice-over artist, who served as the impersonator in our study. She is a middle-aged female whose first language is English. The speech data consists of recordings of a single sentence by the voice-over artist in nine different voices as listed in Table I. The following sentence which is from the VOICES 1.0 database [9] was used: *To further his prestige, he occasionally reads the Wall Street Journal*. Each voice represents an impersonation of a distinct (fictional) character by the voice-over artist, differing from each other in terms of age, gender and regional identity. The voice-over artist used a relatively neutral variety of English (resembling many British influenced Asian standard varieties) for all the voices except NF and SOF, which were perceived by a trained phonetician as having a markedly North American quality. Two distinct voice qualities, nasalization and creak, were also utilized by the voice-over artist for NF and SOF respectively. All the voices were intended to be emotionally neutral, and this impression was confirmed by subsequent inspection.

The recordings were done in a professional home studio using a Shure PG98 Dynamic Vocal microphone with Edirol 24 bit / 96 kHz USB Audio capture and Sonar LE software package for audio processing. The speech data was originally recorded in MP3 format at a sampling rate of 44 kHz using 16 bits/sample. It was later converted to mono WAV format and downsampled to 16 kHz using Audacity®.

III. ANALYSIS AND RESULTS

The speech data was analyzed for three acoustic parameters: fundamental frequency, speech rate and formant frequencies. According to the source-filter theory of speech production [10], the fundamental frequency and speech rate are the source parameters while the formants are the vocal tract parameters. Praat [11] was used to extract the fundamental frequency and the formant frequencies. The fundamental frequency was calculated every 10 milliseconds which is the default step size in Praat while its range was set to 75-600 Hz. The speech rate was calculated using the algorithm in [12] which computes the number of intensity maxima over the entire duration of the speech signal to give an estimate of the number of syllables. The speech rate is thus expressed in terms of the number of syllables per second. The formant frequencies were calculated every 6.25 milliseconds with the maximum formant frequency set to 5500 Hz during the analysis. The number of poles were set to 12. The speech data was segmented at both the word and phone[1] level using the forced alignment algorithm from the HTK Speech Recognition Toolkit [13]. The CMU dictionary [14] was used to provide the phones for each word. The results of the automated segmentation were then manually corrected by a trained phonetician.

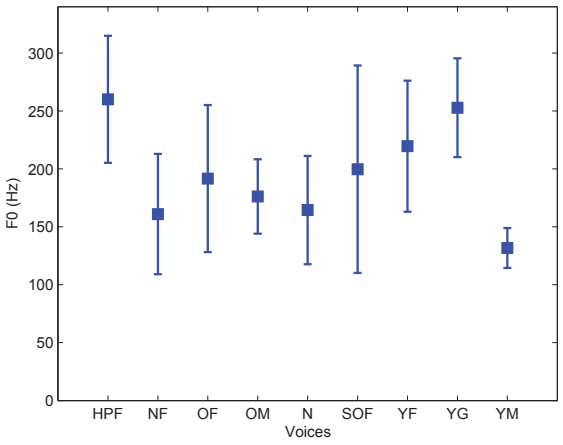

Fig. 1. Mean and standard deviation of F0 for the nine voices calculated using Praat.

A. Fundamental frequency (F0)

Fundamental frequency is the acoustic correlate of perceived pitch in speech. Since F0 characteristics of speech may vary significantly from speaker to speaker, it is important to consider their relevancy for voice identity. The mean fundamental frequency depends largely on the size of the vocal folds. In general, men have lower values of mean F0 compared to women since they have larger vocal folds [15]. It is therefore interesting to see how the impersonator changes her F0 mean and standard deviation with respect to the nine voices. Figure 1 shows the mean and the standard deviation of F0 for all the voices. From the graph, it can be seen that the voice artist is flexible with her mean F0 and can vary it within a range of 132 to 260 Hz. The voice-over artist's natural voice N has a mean F0 of 165 Hz which is rather low for an English speaking female [16].

From Figure 1 it is also observed that the standard deviation of F0 for the male voices i.e. OM and YM is smaller compared

[1] A phone in phonetics is a speech segment which possesses distinct physical and perceptual properties.

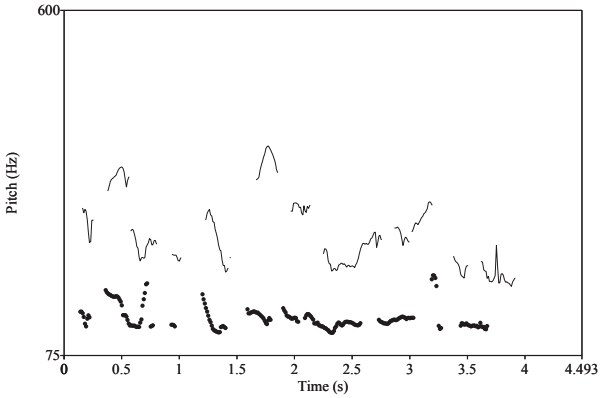

Fig. 2. Pitch contours for HPF (solid line) and YM (bold line) calculated using Praat.

to the female voices. As a result, the male voices sound more monotonous. Since women tend to vary their F0 more than men [17], [18], varying the F0 may be one strategy for feminizing the voice. These differences in the variation of pitch for male and female voices can also be observed from the pitch contour plots. The pitch contour plots represent the evolution of the perceived pitch of the sound over time. Figure 2 shows the pitch contour for HPF and YM. It can be observed that HPF has more peaks and valleys compared to YM, particularly after 1.5 seconds. Additionally, the difference in the peaks and valleys is more pronounced for HPF reflecting a larger F0 range. These differences in the pitch contour clearly show that the voice artist varies her intonation patterns when impersonating male and female voices. Impressionistically, it can be noted that the overall effect is that HPF sounds more expressive than YM.

It is also worth mentioning here that intonation, or variations in pitch height and amplitude that are temporally aligned to the segmental description, is an important parameter for the production of natural sounding speech. However, currently even the state of the art signal processing-based systems [4], [19] simply change the pitch contour of the source speaker $F0_s$ to the pitch contour of the target speaker $F0_t$ according to

$$F0_t = \mu_t + \frac{\sigma_t}{\sigma_s}\,(F0_s - \mu_s) \tag{1}$$

where μ_s, σ_s, μ_t, σ_t represent the mean and standard deviation of the F0 for the source and target speakers respectively.

The linear transformation in Equation (1) fails to capture finer intonational details of the target speaker and to model the local changes in F0. As a result, the converted voice often lacks naturalness specially when converting from male to female voice [4]. Recently some methods [20], [21] have started to include detailed prosody modeling in voice conversion systems but there is still ample room for improvement.

B. Speech Rate

Our findings show that the speech rate is also an important cue for voice identity. It reflects among other factors the speaking style of an individual. The effect of gender and age on speaking rate has been investigated before. It has been shown that men generally speak faster than women [22], [23], [24] while young adults tend to speak faster than older adults [23], [24], [25]. Speech rate was therefore examined in order to understand how the speaker exploits changes in her articulation rate to achieve different voice identities. Figure 3 shows the speech rate for the nine voices. It is observed that the male voices YM and OM have a higher speaking rate compared to the female voices. It can also be observed from Figure 3 that the speech rate of the voice-over artist's natural voice N is closer to all the female voices except NF, YF and YG. The YG voice has the highest speaking rate among all the female voices.

These results suggest that the speech rate is an important parameter used by the voice-over artist for impersonating different ages and genders. On the other hand, voice conversion systems typically employ a time scale modification on a frame by frame basis to adapt the local speaking rate of the source speaker to that of the target speaker. However, the speaking style is lost because of the frame by frame processing [26].

C. Formant frequencies

Formant frequencies are identified by the peaks in the spectral envelope of the speech signal, and are determined by the natural resonances of the vocal tract. For a given speaker, changes in formant frequencies depend primarily on changes in the shape and position of the articulators (tongue, lips, jaw, etc.) during speech production. For linguistic purposes, the first three formant frequencies, F1, F2 and F3, are the principle acoustic correlates of perceptual differences among vowel categories, and are also responsible for subtle differences between vowel tokens within a category. Crucially, formant values also depend inversely on vocal tract length. In general, men have a vocal tract about 20 cm longer than females [27] so it is expected that men have lower overall formant frequencies than females [28] when producing the same vowel. Given that formant frequencies can be an important cue to differences between speakers, they are predicted to be an important feature for voice identity [29]. In order to understand how the voice artist exploits changes in her vocal tract while impersonating, we have analyzed two vowels here, specifically, /ɜ:/ and /ɔ:/. It is found that F1 and F2 of the stressed vowel /ɜ:/ in *journal* and the stressed monophthong /ɔ:/ in *wall* clearly demonstrate the strategies employed by the voice artist to change her identity. This is illustrated in Figures 4 and 5 which show the F1 vs F2 plot for the two vowels /ɜ:/ and /ɔ:/ respectively. It is observed that the male voices: OM and YM have a lower value of F1 compared to the female voices. The natural voice of the voice-over artist N is closer to the SOF and OF voices while the young female YF has a higher F1. However, the NF and HPF voices do seem to change position with the vowel. The YG has a particularly high value of F2. These results suggest that the voice-over artist is able to alter her perceived vocal tract length while impersonating different ages, genders and voice qualities.

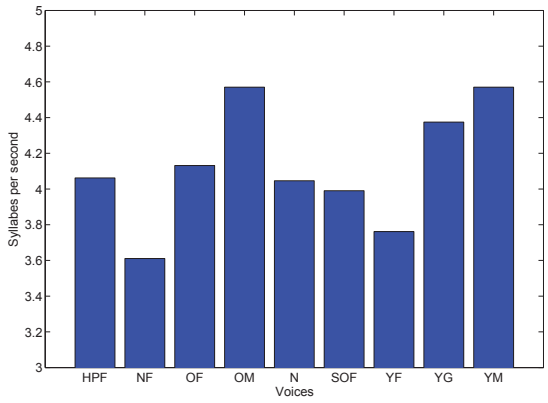

Fig. 3. The speech rate for the nine voices in terms of syllables per second calculated using [12].

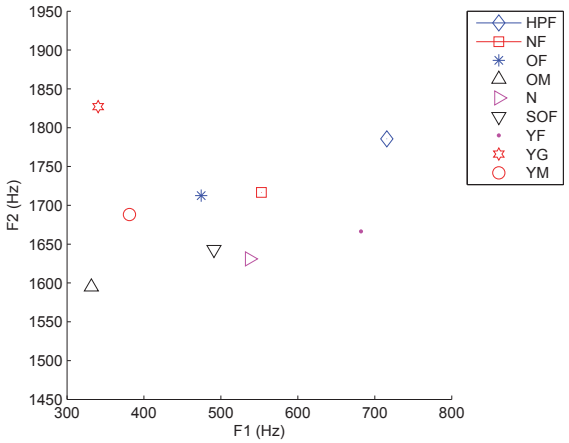

Fig. 4. F1 vs F2 for /ɜ:/ in *journal* for the nine voices.

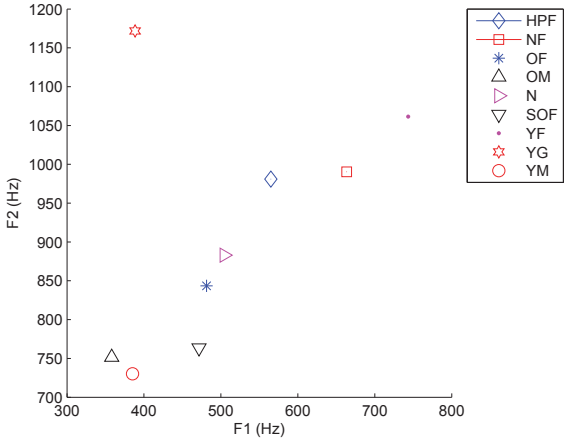

Fig. 5. F1 vs F2 for /ɔ:/ in *wall* for the nine voices.

Linguistically, these two vowels are somewhat special since they are essentially key markers of regional accent. For example, /ɜ:/ is pronounced with /r/-coloring, or rhotacization, in only some of the voices, namely NF and SOF. The primary acoustic correlate of such rhotacization is a lowering of F3, though it may also be accompanied by slight raising of F2 [30]. This effect can be observed in Figure 6 which shows the first three formant frequencies for /ɜ:/ in *journal*. Perceptually, the overall impression of F3-lowering is that the speaker is using a North American dialect of English. At least part of the discriminability seen in Figure 4, therefore, may be a result of the speaker's ability to access different regional accents, rather than from voice quality per se. The other vowel /ɔ:/ is similar, in that the degree of lip rounding, manifested acoustically as a lowering of F2, is highly associated with differences in regional accent. Still, in the case of /ɔ:/, it is possible that the voice-over artist may have been exploiting changes in F2 via lip rounding to achieve speaker-specific effects that are unrelated to regional accent such as vocal tract length or cross-sectional area.

Together, these results show that the high degree of variability exhibited in the voice-over artist's production of vowels is clearly a major resource that she exploits to achieve different voice identities.

IV. Conclusions

In this paper, we have performed an acoustic and linguistic analysis of nine different voices produced by a voice-over artist. From auditory analysis, it is clear that she is successful in changing her voice identity in terms of age, gender and voice quality. The results also confirm that these changes are accompanied by changes in mean F0, F0 dispersion, intonation pattern, speaking rate and vocal tract shape to achieve different voice identities. Furthermore, we gave full consideration to voice identity features. It is concluded that the acoustical measures are greatly affected by language-dependent features such as those associated with regional accent.

These findings suggest that it is possible to synthetically generate a range of distinct voice identities by systematically recombining a finite number of linguistic and non-linguistic features of the speech signal. Further research is needed to determine whether the set of factors considered in our study are sufficient for generating voices that rate highly in terms of individuality and naturalness, or whether more features must be incorporated. This would be useful for developing speech synthesis algorithms which are capable of producing a range of different natural voices from a single voice. To further generalize our observations, our future work involves extending the analysis to a larger set of sentences and paragraphs from impersonators of different ages, genders, accents and dialects.

Acknowledgments

We would like to thank the voice-over artist Noella Menon from *voice4ads* for the voice impersonations. The first author would also like to acknowledge Samuel Deslauriers-Gauthier for his inputs on the implementation of the acoustic analysis interface.

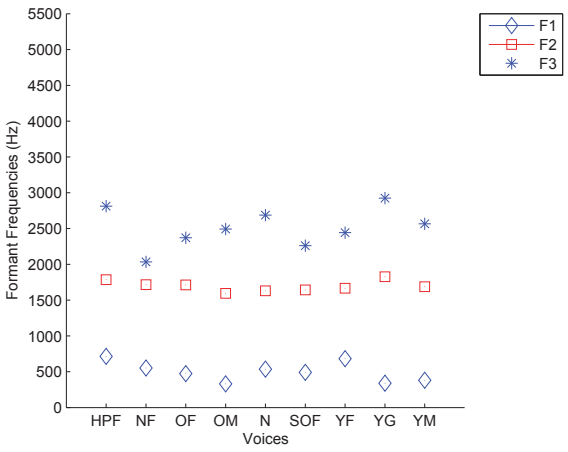

Fig. 6. The first three formants: F1, F2, F3 for /ɜ:/ in *journal* for the nine voices.

REFERENCES

[1] H.F. Hollien, *Forensic voice identification*, Academic Press, 2002.

[2] J.F. Bonastre, F. Bimbot, L.J. Boë, J.P. Campbell, D.A. Reynolds, and I. Magrin-Chagnolleau, "Person authentication by voice: A need for caution," in *Eighth European Conference on Speech Communication and Technology*, 2003, pp. 33–36.

[3] Y. Stylianou, "Voice transformation: A survey," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3585–3588.

[4] A.F. Machado and M. Queiroz, "Voice conversion: A critical survey," *Proc. Sound and Music Computing (SMC)*, 2010.

[5] A. Eriksson and P. Wretling, "How flexible is the human voice?-a case study of mimicry," in *Fifth European Conference on Speech Communication and Technology*, 1997, pp. 1043–1046.

[6] E. Zetterholm, "Same speaker–different voices. a study of one impersonator and some of his different imitations," in *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, 2006, pp. 70–75.

[7] T. Kitamura, "Acoustic analysis of imitated voice produced by a professional impersonator," in *Proc. Interspeech*, 2008, pp. 813–816.

[8] E. Zetterholm, "Impersonation: a phonetic case study of the imitation of a voice," *Lund Working Papers in Linguistics*, vol. 46, pp. 269–287, 2009.

[9] A. Kain, *High resolution voice transformation*, Ph.D. thesis, OGI School of Science and Engineering, Oregon Health and Science University, 2001.

[10] G. Fant, *Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations*, Mouton & Co. N.V., The Hague, 1970.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 5.3," Retrieved October 21, 2011, from http://www.praat.org/.

[12] N.H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK Version 3.4)," *Cambridge University Engineering Department*, 2006.

[14] R.L. Weide, "The CMU pronunciation dictionary, Release 0.6d," 1998, ftp://ftp.cs.cmu.edu/project/speech/dict/.

[15] M. Latinus and P. Belin, "Human voice perception," *Current Biology*, vol. 21, no. 4, pp. R143–R145, 2011.

[16] WS Brown, R.J. Morris, H. Hollien, E. Howell, et al., "Speaking fundamental frequency characteristics as a function of age and professional singing," *Journal of Voice*, vol. 5, no. 4, pp. 310–315, 1991.

[17] J. van Rie and R. van Bezooijen, "Perceptual characteristics of voice quality in dutch males and females from 9 to 85 years," in *Proceedings of the XIIIth International Congress of Phonetic Sciences 2*, 1995, pp. 290–293.

[18] R.M. Brend, "Male-female intonation patterns in american english," *Language and sex: Difference and dominance*, vol. 86, 1975.

[19] D. Erro, A. Moreno, and A. Bonafonte, "Flexible harmonic/stochastic speech synthesis," in *6th ISCA Workshop on Speech Synthesis*, 2007.

[20] E.E. Helander and J. Nurminen, "A novel method for prosody prediction in voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, vol. 4, pp. 509–512.

[21] Z. Inanoglu, "Transforming pitch in a voice conversion framework," M.S. thesis, St. Edmonds College, University of Cambridge, 2003.

[22] D. Byrd, "Relations of sex and dialect to reduction," *Speech Communication*, vol. 15, no. 1-2, pp. 39–54, 1994.

[23] J. Yuan, M. Liberman, and C. Cieri, "Towards an integrated understanding of speaking rate in conversation," in *Ninth International Conference on Spoken Language Processing*, 2006, pp. 541–544.

[24] E. Jacewicz, R.A. Fox, C. O'Neill, and J. Salmons, "Articulation rate across dialect, age, and gender," *Language variation and change*, vol. 21, no. 02, pp. 233–256, 2009.

[25] B.L. Smith, J. Wasowicz, and J. Preston, "Temporal characteristics of the speech of normal elderly adults," *Journal of Speech and Hearing Research*, vol. 30, no. 4, pp. 522–529, 1987.

[26] L. Leutelt and U. Heute, "Voice conversion: Adaptation of relative local speech rate by MPEG-4 HVXC," in *EUSIPCO*, 2002, vol. 3, pp. 113–116.

[27] G. Fant, "A note on vocal tract size factors and non-uniform f-pattern scalings," *Speech Transmission Laboratory Quarterly Progress and Status Report*, vol. 1, pp. 22–30, 1966.

[28] G.E. Peterson and H.L. Barney, "Control methods used in a study of the vowels," *Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[29] R.O. Coleman, "A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice," *Journal of Speech and Hearing Research*, vol. 19, no. 1, pp. 168–180, 1976.

[30] K.N. Stevens, *Acoustic phonetics*, vol. 30, The MIT press, 2000.