Preliminary Evidence of Sexual Bias in Voice over Internet Protocol Audio Compression

Matthew L. Bolton(✉)

University of Virginia, Charlottesville, VA 22903, USA
mlb4b@virginia.edu

Abstract. Voice over internet protocol (VoIP) has become critical to professional communication through its use in office phone systems and video teleconferencing. In this environment, it is critical that VoIP represent everybody's voices with equal fidelity. This will allow all demographics to have the opportunity for equal participation. This research investigated whether there was evidence of sexism in the way VoIP compresses vocal data. This was accomplished by measuring the similarity of a large database of vocal samples between their original and compressed forms and statistically comparing their differences between females and males. Multiple measure showed significant evidence of reduced audio quality for female voices compared to male voices. We discuss these results and propose future research.

Keywords: Access to on-line communities and eServices · Universal access and accessibility requirements · Voice communication · Audio compression

1 Introduction

Even before the COVID-19 pandemic, voice over internet protocol (VoIP) was important for people conducting business. This is because it allows voice communication (with or without video) to occur with nearly anybody in the world over the internet. Since the COVID-19 outbreak and the associated social distancing, VoIP has become critical for enabling remote operations. Even when this situation resolves, VoIP will remain an integral part of modern work. It is thus critical that the understandability and quality of VoIP presented speech be equal for all demographics of people. However, in my own subjective experience using VoIP to communicate with my wife, students, and colleagues, I have had more trouble understanding female voices (for people I had no trouble understanding in person) over VoIP than male voices.

This research sought to be a preliminary investigation into whether there is evidence of sexist performance in VoIP communication. Below, we present background for understanding the presented approach, the study's scientific objectives, methods, and results. Ultimately, findings are discussed and future research directions are identified.

M. Kurosu (Ed.): HCII 2022, LNCS 13303, pp. 227–237, 2022.
https://doi.org/10.1007/978-3-031-05409-9_17

2 Background

To the best of my knowledge, no study has investigated whether there are differences in how VoIP handles female and male voices. Thus, no literature on this subject is reviewed. However, an understanding of VoIP and "lossy" audio codec (compression decompression) technology is topical to the research. These are discussed below.

In VoIP, a voice is processed by a microphone, encoded into a digital format, transmitted over the internet to a recipient device, and decoded and played on a speaker. Audio codecs are critical to the performance of VoIP's encoding process. Audio codecs constitute methods for compressing audio files so that they can be transmitted efficiently. While lossless codec's exist (codecs that loose no audio data), most VoIP system codecs are lossy. That is, they eliminate audio data to reduce the amount of information that needs to be transmitted. The way that codecs do this can range from using filtering techniques to more sophisticated approaches that use psychoacoustics to eliminate frequencies that are simultaneously masked (rendered inaudible) by other concurrent sounds due to limitations of the human sensory system [1]. Most codecs can also support varying bitrates, a measure of how many bits are used to represent a second of audio data. The lower the bitrate, the smaller the file. When used in many modern VoIP systems [5], the bandwidth or ping of the connection will be used to dynamically vary the bitrate to ensure the audio is delivered continuously.

There are a number of different codecs used in VoIP such as G.711, G.719, OPUS, G.722, G.729 and AMR. For the purpose of this work, I focus on OPUS [17]. This is because OPUS is the audio codec used in a number of popular VoIP applications including ZOOM, Skype, Microsoft Teams, and WhatsApp. It is also extremely flexible as it is low latency, offers sophisticated lossy audio compression specifically tuned for speech, and can use bitrates ranging from 6 to 510 kbps that can be varied based on bandwidth availability [5,17].

3 Objective

This work sought to determine if there was evidence of sexism in VoIP. To accomplish this we made use of an online database of English speaking for males and females with varying ages and native languages. These were each processed in accordance with minimal quality settings using the OPUS audio codec. These settings were used because they would constitute the largest possible distortion of the original signal. The *original* and resulting *compressed* files were compared across a number of standard auditory similarity measures. These measures were then statistically compared between male and female voices to understand which was the most distorted or degraded and identify the nature of the change.

4 Methods

4.1 Data and Apparatus

The data used in this experiment constituted the full set of audio samples (files) from the Speech Accent Archive [19,20]. This is a resource available online through a Creative Commons License. The archive contains 2,933 separate audio files (when pruned for blank entries and computer generated voices), where each is a speech sample from a unique person. The samples were taken from people of different sexes, ages, and language backgrounds. The sample ultimately included 1494 females and 1439 males with an age range of [6] with a median of 27. In every audio sample, the speaker recites the same paragraph in English:

> Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

All the samples were encoded in monaural MP3 with a 48,000 Hz sample rate and 256 kbps bitrate. Note that MP3 is itself a lossy audio codec. However, the sample rate and bitrate of these files are well above thresholds that would influence audio quality perception in normal listeners.

To simulate the effect of the audio being broadcast over VoIP using a slow internet connection, each *original* MP3 was *compressed* using the OPUS audio codec using the lowest possible settings: a constant bitrate of 6 kbps. The option for optimizing for speech was also selected.

4.2 Independent Variables

In all the presented analyses, *Sex*, with Female and Male levels, was the independent variable.

4.3 Dependent Measures

In this research, metrics were selected that would measure the similarity between an individual's *original* and *compressed* audio samples. There are many metrics for accomplishing this [12]. In this work, ten metrics were selected that appeared to have the most precedent in the literature. All are described below along with how they were computed.

Cross Correlation. The cross correlation was a measure of the correlation of raw waveform data of *original* and *compressed*. This was computed as the value returned for zero lag by MATLAB's `xcorr` function. The higher the produced correlation the more similar *original* is to *compressed*.

Signal-to-Noise Ratio (in dB). Signal-to-noise ratio was computed with MATLAB's `snr` function: the ratio of the power of *compressed* ("signal") to that of *original* ("noise") in dB. The lower the value, the more power lost in *compressed*.

Euclidian Distance. The Euclidian distance between *original* and *compressed* was computed as the norm (or vector magnitude) of the vector formed by *original* − *compressed*. This was calculated using MATLAB's `norm` function. The smaller the distance, the more similar the audio samples.

Mean Coherence. Mean coherence was computed as the average of the magnitude-squared coherence of *original* and *compressed* with a window size of 25 ms, an overlap of 12.5 ms, and 256 sample points, all parameters for the required Fourier transform. These parameters were selected because they are the standard for analyzing speech audio data [11]. This was computed as the average of the vector returned by MATLAB's `mscohere` function. The higher the coherence, the more similar the two audio signals.

Compression Error Rate. The compression error rate was introduced by Siegert et al. [13] as a measure for quantifying data loss from lossy audio codecs. This was computed using the formulation from [13] as the absolute error obtained by comparing the spectrograms of *original* and *compressed*, using the standard Fourier transform parameters defined above. The lower the error rate, the more similar the audio samples.

Spectral Entropy Correlation. Spectrally entropy is a measure of a signal's irregularity [8]. Thus, spectral entropy correlation represents a measure of the correspondence between the signal irregularities of *original* and *compressed*. This was computed as the correlation between the vectors returned by MATLAB's `pentropy` function. The higher the correlation, the more similar the samples.

Structural Similarity. Structural similarity is commonly used to measure the similarity between original on compressed images [18]. However, it can also be used for similar purposes for audio [3,7]. Structural similarly was computed here using MATLAB's `ssim` function. The higher the structural similarity score, the more similar the two audio signals.

Objective Difference Grade. The objective different grade (ODG) is meant as an objective measure of perceived audio quality. It is computed in accordance with the ITU BS.1387-1. (PEAQ) algorithm [15]. This was designed to simulate the human ear and incorporate psychophysical concepts to produce a single metric of perceived audio quality when comparing an original signal to a degraded or compressed one. This metric ranges from 0 (imperceptible difference) to −4

(annoying). This metric, and the algorithm for computing it, are an international standard and are thus used in a number of scientific studies and the development of many audio technologies. In this work, we used the MATLAB ODG implementation called `PQevalAudio` provided by Kabal [6].

Centroid Difference (in Hz). The spectral centroid is a measure for characterizing the mean or average of a signal's frequencies [12]. It has been experimentally shown to have a robust relationship with the perceived brightness of sounds [4]. It is ultimately calculated as the weighted average of a signal's frequencies based on a Fourier transform [12]. In this work, the centroid difference is calculated as the spectral centroid of *original* minus the spectral centroid of *compressed*, both computed using MATLAB's `spectralCentroid` function with the Fourier transform parameters from above. Thus, centroid difference represents a measure of how the frequency or perceived brightness of the sound changed when the sound was compressed. The larger the magnitude of the centroid difference, the more of a shift occurred. A positive results means *compressed* is less bright (favors lower frequencies) than *original*. A negative results means *compressed* is brighter (favors high frequencies) than *original*.

Spread Difference (in Hz). The spectral spread of a sound is a measure of the variance of a signal's frequencies, the "spread" of the sound around its spectral centroid [12]. Noisy sounds usually exhibit a large spectral spread. Tonal sounds usually have lower spectral spreads. In this research, the spread difference is calculated as the spectral spread of *original* minus the spectral spread of *compressed*, where spreads are computed with MATLAB's `spectralSpread` function with the Fourier transform parameters from above. The larger the spread difference, the more spectral variation is lost. A positive value implies that *compressed* has less variance than *original*. A negative value implies the opposite.

4.4 Data Analysis

A two-sample Mann-Whitney U Test was used to compare differences between each of the computed measures for samples taken from female and male speakers. This nonparametric test was used due to the failure of the dependent measures to show evidence of normal distribution using Anderson Darling tests. A significance level of 0.05 was used. However, to compensate for multiple tests (one for all ten dependent measures), we employed a Bonferoni correction that lowered the significance level to $0.05/10 = 0.005$.

5 Results

Results (Table 1) revealed significant differences between females and males for signal-to-noise ratio, Euclidian distance, compression error rate, spectral entropy correlation, objective difference grade, and spread difference.

Table 1. Results comparing females and males for each dependent measure

Measure	M_{Females}	M_{Males}	$U_{1494,1439}$	z	p	r
Cross correlation	0.791	0.783	1032591.0	−1.850	0.065	0.034
Signal-to-noise ratio (dB)	−1.373	−1.718	653834.0	−18.400	< 0.001 *	0.340
Euclidian distance	48.843	45.958	930913.0	−6.280	< 0.001 *	0.116
Mean coherence	0.195	0.191	1074740.5	−0.008	0.993	< 0.001
Compression error rate	0.195	0.183	989949.0	−3.710	< 0.001 *	0.069
Spectral entropy correlation	0.784	0.827	1270719.0	8.540	< 0.001 *	−0.158
Structural similarity	0.699	0.697	1048410.0	−1.160	0.247	0.021
Objective difference grade	−3.913	−3.913	1198438.5	6.050	< 0.001 *	−0.112
Centroid difference (Hz)	122.770	99.207	1013949.0	−2.660	0.008	0.049
Spread difference (Hz)	579.573	435.738	957807.0	−5.110	< 0.001 *	0.094

Note. M_{Females} and M_{Males} represent the median values for females and males respectively for each measure. U is the Mann-Whitney U test statistics with sample sizes $n_{\text{Female}} = 1494$ and $n_{\text{Male}} = 1439$. z is the associated z-score test statistic. p is the test's p-value. A p-value is labeled with a * if it was significant at a 0.005 level. r, rank-biserial correlation [2], is the effect size.

A closer comparison of the results (Fig. 1) revealed that, for all but signal-to-noise ratio (where the higher value observed for females indicates more correspondence between *original* and *compressed*), the *compressed* sound resulted in more of a change for female voices than for males. For female samples, there was a significantly larger Euclidian distance between *original* and *compressed*, a significantly larger compression error rate, a significantly lower spectral entropy correlation, a significantly lower objective difference grade, and a significantly higher spread difference. It should be noted that objective difference grade (Fig. 1(E)) showed very little variance in the results: $[-3.9132, -3.837]$ for females and $[-3.9132, -3.8275]$ for males. In fact, 1729 of the computations (58.95%) exhibited the minimum value of -3.9132.

6 Discussion

The results show clear evidence of bias in the OPUS VoIP audio codec at the low bitrate used in the analysis. Given that five of the six significant measures showed a bias that favored male over female voices, this suggest that the OPUS codec is likely, unintentionally, sexist towards women. This is a potentially serious issue given that OPUS is used as the primary audio codec of ZOOM, Skype, Microsoft Teams, WhatsApp, Playstation Network, and many VoIP phone service vendors and it may be limiting women's ability to communicate.

A possible explanation for the observed differences is illustrated in Fig. 2. This compares the spectrum of an *original* and *compressed* pair for a female speaker. While there are clear differences, the biggest occurs above 4,500 Hz. Specifically, the *compressed* sample contains no frequencies greater than this value, while the *original* clearly does. This suggests that OPUS (at least with the selected parameters) uses a low pass filter to eliminate audio data. Given that women tend to have higher pitched voices than men [16], this would likely impact females more

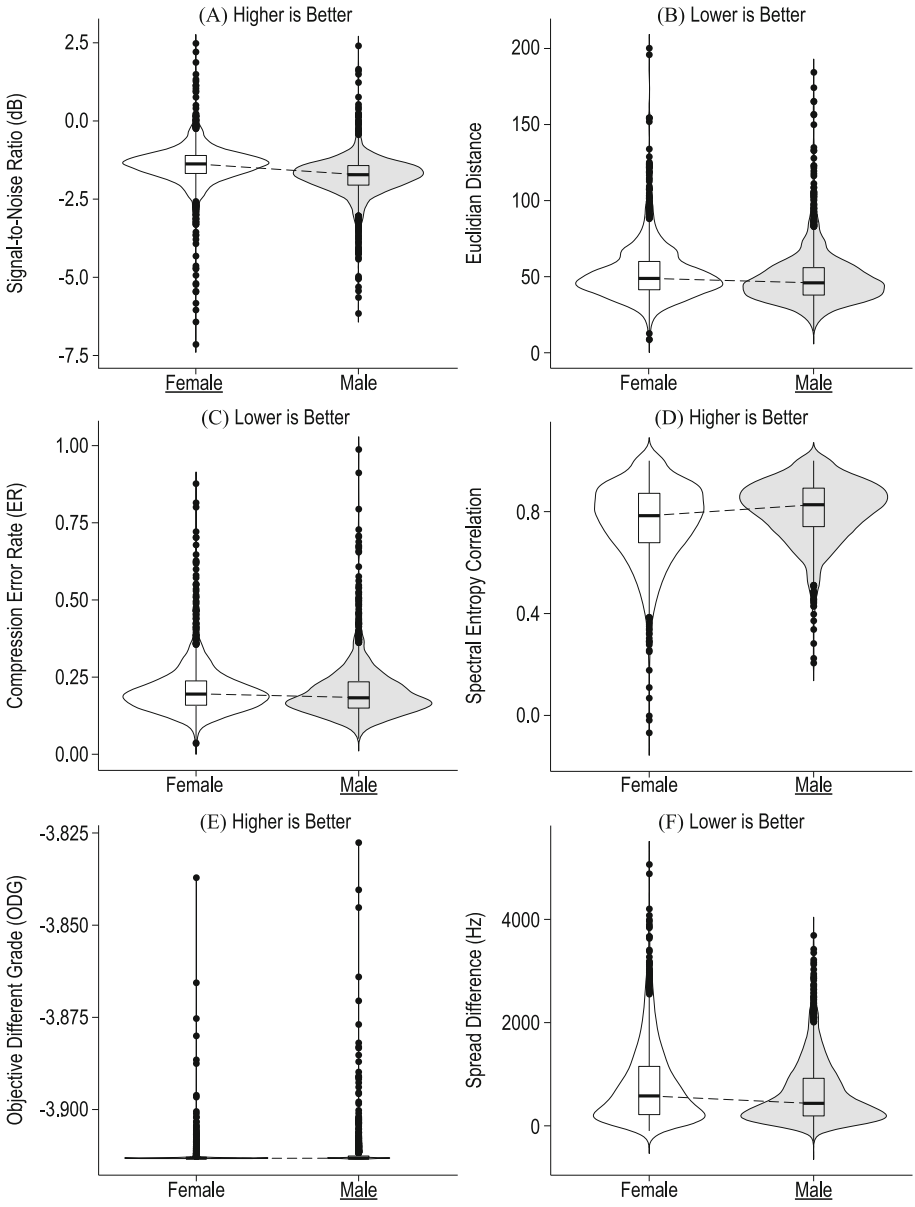

Fig. 1. Violin plots for dependent measures that showed significant differences between Females and Males. Each entry contains a box and whisker plot surround by kernel density plots. The box depicts the interquartile range (IQR) bisected horizontally by a line at the median (M); whiskers represent the extremum within 1.5 of the IQR; outliers (filled dots) are points outside of 1.5 of the IQR. A dotted line between plot medians illustrates changes between the values between levels. Underlined x-axis labels indicate which category (Female or Male) saw better median performance.

than males. This is reinforced by the results and, specifically, the bigger median observed for spread difference for female voices than male ones. Although not statistically significant with our adjusted level ($p = 0.008$), this is also consistent with the results seen for centroid difference, where we saw a bigger downward shift in pitch (centroid difference) for female voices ($M_{\text{Female}} = 122.770\,\text{Hz}$) than male ones ($M_{\text{Male}} = 99.207\,\text{Hz}$). This view also potentially explains the outlier result seen for signal-to-noise ratio. Specifically, signal-to-noise ratio is based on the correspondence between power (the intensity or volume) of the signals. Figure 2 also suggests that the volume/power of lower frequencies are reduced in compressed files compared to the included higher frequencies. Given that male voice favor lower frequencies, the reduction of the volume of these frequencies could explain their lower signal-to-noise ratio.

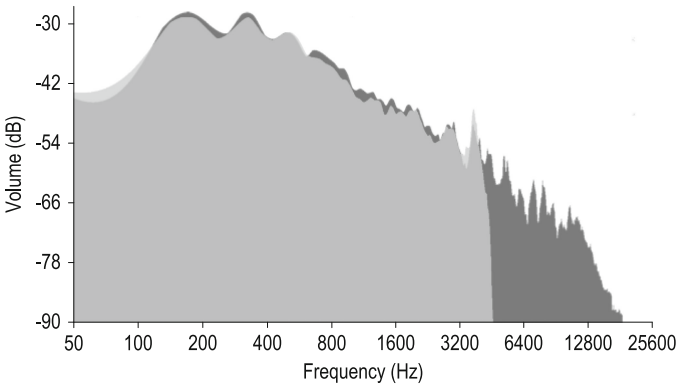

Fig. 2. Comparison of spectrum plots for a random *original* (dark gray) and *compressed* (lite gray) sample pair, where the spectrum for *compressed* is semi-transparent and overlaid on the spectrum of *original*.

Based on the history of voice communication technology, the results make sense. As surveyed by Monson et al. [9], vocal telecommunication has generally focused on frequencies below 5000 Hz because, due to technological limitations, historical evidence showed these were most important. However, there is now good evidence that higher frequencies sounds are important to speech communication. Monson et al. [9] found that humans could detect differences in levels in high frequency sounds, especially in speech. Moore and Tan [10] found that higher frequencies were important for the perceived quality of voice audio. Finally, Snow [14] found significant increases in speech discrimination when low pass filter cutoffs were increased. This was especially true for female voices, which required higher cutoffs (well above 5000 Hz) for comparable discrimination rates with male voices. All of this provides more evidence that the loss of higher frequencies in VoIP is disadvantageous to effective and natural communication for women.

The effect sizes observed in the results are small. However, the above discussion shows that even these small difference may impact an individual's ability to communicate. Beyond this, the scale at which VoIP technology like OPUS is used suggests that, as marginal as individual impacts could be, there is very likely an aggregate detrimental impact on female users.

The results presented here are preliminary. There are many avenues for future investigation. These are discussed below.

6.1 Objective Difference Grade

The extremely small range of values obtained for the objective difference grade measure suggest that the OPUS codec is specifically designed to achieve specific levels of objective difference grade performance. Thus, even though significant difference were observed between male and female voices, there was not a practically significant difference between median scores. This suggests that the objective difference grade may not be the most appropriate metric for determining how to scale speech. Future work should investigate if there are other measure that would better suited for understand the quality of speech signals while accounting for the sex of the speaker.

6.2 Other Settings, Codecs, and VoIP Considerations

The analysis presented here is far from complete. The OPUS codec offers bitrates all the way up to 512 kbps, which will likely reduce differences between signals significantly. Additionally, as discussed earlier, there are other common codecs used in VoIP applications, each with their own variable settings. Beyond this, there may be other factors that could impact speech comprehension and quality in VoIP such as the rate of packets being dropped and the amount of information loss this constitutes in the audio signal. Future work should investigate to what extent the biases reported here manifest with other settings, codecs, and VoIP considerations.

6.3 Additional Voice Considerations

Beyond sex, there are other factors that could influence how humans use different frequencies in voice communications. This could suggest that their may be biases against people with different accents or communicating in different languages. This should be explored in future efforts.

6.4 Experimental Validation

While the results presented here are compelling and there is literature suggesting significant real world impact, the magnitude of this impact is not entirely clear. Thus, future work should conduct an experiment with human subjects to validate our results and assess real word effects.

References

1. Bosi, M., Goldberg, R.E.: Introduction to Digital Audio Coding and Standards. Springer, Heidelberg (2012)
2. Cureton, E.E.: Rank-biserial correlation. Psychometrika **21**(3), 287–290 (1956)
3. Gan, C., Wang, X., Zhu, M., Yu, X.: Audio quality evaluation using frequency structural similarity measure. In: IET International Communication Conference on Wireless Mobile and Computing (CCWMC 2011), pp. 299–303. IET (2011). https://doi.org/10.1049/cp.2011.0896
4. Grey, J.M., Gordon, J.W.: Perceptual effects of spectral modifications on musical timbres. J. Acoust. Soc. Am. **63**(5), 1493–1500 (1978)
5. Han, Y., Magoni, D., McDonagh, P., Murphy, L.: Determination of bit-rate adaptation thresholds for the OPUS codec for VoIP services. In: 2014 IEEE Symposium on Computers and Communications (ISCC), pp. 1–7. IEEE (2014)
6. Kabal, P.: An examination and interpretation of ITU-R BS. 1387: perceptual evaluation of audio quality. Technical report, Telecommunications and Signal Processing Laboratory, McGill University, Montreal (2002)
7. Kandadai, S., Hardin, J., Creusere, C.D.: Audio quality assessment using the mean structural similarity measure. In: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 221–224. IEEE (2008)
8. Llanos, F., Alexander, J.M., Stilp, C.E., Kluender, K.R.: Power spectral entropy as an information-theoretic correlate of manner of articulation in American English. J. Acoust. Soc. Am. **141**(2), EL127–EL133 (2017)
9. Monson, B.B., Hunter, E.J., Lotto, A.J., Story, B.H.: The perceptual significance of high-frequency energy in the human voice. Front. Psychol. **5**(587) (2014). 11 pages
10. Moore, B.C., Tan, C.T.: Perceived naturalness of spectrally distorted speech and music. J. Acoust. Soc. Am. **114**(1), 408–419 (2003)
11. Paliwal, K.K., Lyons, J.G., Wójcicki, K.K.: Preference for 20–40 ms window duration in speech analysis. In: 2010 4th International Conference on Signal Processing and Communication Systems, pp. 1–4. IEEE (2010)
12. Peeters, G.: A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical report 23/04/04, Ircam, Analysis/Synthesis Team, Paris (2004)
13. Siegert, I., Lotz, A.F., Duong, L.L., Wendemuth, A.: Measuring the impact of audio compression on the spectral quality of speech data. In: Jokisch, O. (ed.) Elektronische Sprachsignalverarbeitung (ESSV 2016). Proceedings of the 27th ESSV Conference (Studientexte zur Sprachkommunikation, vol. 81), pp. 229–236. TUDpress, Dresden (2016)
14. Snow, W.B.: Audible frequency ranges of music, speech and noise. Bell Syst. Tech. J. **10**(4), 616–627 (1931)
15. Thiede, T., et al.: PEAQ-the ITU standard for objective measurement of perceived audio quality. J. Audio Eng. Soc. **48**(1/2), 3–29 (2000)
16. Titze, I.R.: Physiologic and acoustic differences between male and female voices. J. Acoust. Soc. Am. **85**(4), 1699–1707 (1989)
17. Vos, K., Sørensen, K.V., Jensen, S.S., Valin, J.M.: Voice coding with opus. In: Audio Engineering Society Convention 135. Audio Engineering Society (2013)
18. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)

19. Weinberger, S.H.: Speech accent archive (2015). George Mason University. http://accent.gmu.edu
20. Weinberger, S.H., Kunath, S.A.: The speech accent archive: towards a typology of English accents. In: Corpus-Based Studies in Language Use, Language Learning, and Language Documentation, pp. 265–281. Brill Rodopi (2011)