Research Note

The Effects of Speech Compression Algorithms on the Intelligibility of Two Individuals With Dysarthric Speech

Rene L. Utianski,^a Steven Sandoval,^b Visar Berisha,^{c,d}
Kaitlin L. Lansford,^e and Julie M. Liss^d

Purpose: Telemedicine, used to offset disparities in access to speech-language therapy, relies on technology that utilizes compression algorithms to transmit signals efficiently. These algorithms have been thoroughly evaluated on healthy speech; however, the effects of compression algorithms on the intelligibility of disordered speech have not been adequately explored.

Method: This case study assessed acoustic and perceptual effects of resampling and speech compression (i.e., transcoding) on the speech of 2 individuals with dysarthria. Forced-choice vowel identification and transcription tasks were utilized, completed by 20 naive undergraduate listeners.

Results: Results showed relative improvements and decrements in intelligibility, on various measures, based

on the speakers' acoustic profiles. The transcoding of the speech compression algorithm resulted in an enlarged vowel space area and associated improvements in vowel identification for 1 speaker and a smaller vowel space area and decreased vowel identification for the other speaker. Interestingly, there was an overall decrease in intelligibility in the transcription task in this condition for both speakers.

Conclusions: There is a complex interplay between dysarthria and compression algorithms that warrants further exploration. The findings suggest that it is critical to be mindful of apparent changes in intelligibility secondary to compression algorithms necessary for practicing telemedicine.

Supplemental Material: <https://doi.org/10.23641/asha.7291940>

Current applications of telemedicine in speech-language pathology rely on existing technology for the compression and transmission of the speech signal developed for cellular and Internet communications. A critical component in these systems is the algorithm for compressing the acquired signal. These algorithms are often utilized in consumer mobile applications (Guillemin & Watson, 2008; Sauter, 2010), including, but not limited to, videoconferencing systems and cell phone transmission,

recording, and playback applications. Past research has demonstrated that such coding schemes preserve both the quality and intelligibility of healthy speech (Barreto & Pirhonen, 2006; Holma, Melero, Vainio, Halonen, & Makinen, 2003; Thorpe, 1999), although there is a common assumption throughout the field that there are some reductions in vocal quality. A basic understanding of these algorithms, as well as their impact on healthy and disordered speech, is critical before their application in telemedicine becomes ubiquitous. Identifying possible effects of the compression algorithms on disordered speech will help speech-language pathologists (SLPs) engaged in telepractice avoid misattribution of changes in speech to the patient when they are possibly transmission artifacts.

The majority of speech compression algorithms are based on linear prediction. Linear prediction assumes that each speech sample can be predicted by preceding samples. At the source, the prediction coefficients and the residuals are encoded using codebooks trained on healthy speech. Codebooks consist of different patterns that capture important aspects of the speech on which they were trained. For example, a codebook may contain a comprehensive set of

^aDepartment of Neurology, Mayo Clinic, Rochester, MN

^bDepartment of Electrical and Computer Engineering, New Mexico State University, Las Cruces

^cSchool of Electrical, Computer and Energy Engineering, Arizona State University, Tempe

^dDepartment of Speech and Hearing Science, Arizona State University, Tempe

^eSchool of Communication Science & Disorders, Florida State University, Tallahassee

Correspondence to Rene L. Utianski: Utianski.Rene@mayo.edu

Editor-in-Chief: Julie Barkmeier-Kraemer

Editor: Kristie Spencer

Received April 23, 2018

Revision received June 8, 2018

Accepted July 10, 2018

https://doi.org/10.1044/2018_AJSLP-18-0081

Disclosure: Visar Berisha is the Chief Science Officer at Aural Analytics. Julie M. Liss is the Chief Clinical Officer at Aural Analytics. The authors have no other relevant conflicts of interest to disclose.

speech spectra that are effectively proxies for the different vowels and consonants in the English language. For bandwidth efficiency, input speech is represented by entries in these codebooks, rather than the raw speech samples. At the destination, the codebook entries are used to reconstruct an estimate of the original speech signal. However, there is good reason to suspect that compression of an already noisy and distorted acoustic signal produced by a speaker with dysarthria may have unintended effects because the codebooks associated with common compression schemes have been trained on large corpora of healthy speech. One prior study looked at the impact of compression algorithms on a variety of voice disorders, including dysarthria, utilizing consonant and vowel accuracy of nonsense words as the outcome measure (Jamieson, Parsa, Price, & Till, 2002). This study demonstrated reductions in intelligibility associated with the number of compression algorithms tested and also noted that these coders interact with individual voices. It has yet to be determined what effect the codebooks trained on healthy speech would have on both the intelligibility and acoustics of connected speech produced by individuals with different types of dysarthria.

Exploring the effects of sampling and encoding manipulations on disordered speech is timely because of the use of these algorithms in the growing practice of telemedicine in speech-language pathology assessment and intervention (American Speech-Language-Hearing Association, 2012). For a remote clinical platform to be successful, the transmission of the speech signal must faithfully preserve the disordered aspects of the speech, including acoustic features of abnormal voice, resonance, articulation, and prosody. As resampling is a shared first step in compression schemes, down-sampled speech is an important comparison to parse the relative contributions of the sampling and subsequent compression. In this report, we examine the effects of sampling rate and compression on speech intelligibility and vowel identification of speech produced by individuals with dysarthria. We selected vowels for this first study because accurate vowel identification supports two fundamental speech perception processes—word recognition and lexical segmentation (Spitzer, Liss, & Mattys, 2007). Inaccurate word recognition and/or lexical segmentation result in reduced intelligibility (Liss et al., 2009). A second reason for selecting vowels for this case study is that there is potential for algorithms to perform differently when the necessary spectral information, particularly formant frequencies, is distorted.

Given the aforementioned information, we aimed to evaluate the impact of a commonly used compression algorithm on dysarthric vowel acoustics and the corresponding perception of this speech. Specifically, we examined the acoustic and perceptual effects of the following conditions: (a) *original*: unaltered and recorded with a 44.1-kHz sampling rate; (b) *resampled*: low-pass filtered at 4 kHz and sampled at a rate of 8 kHz, as is standard in cellular networks; and (c) *transcoded*: processed with a compression algorithm known as the *Global System for Mobile communications Adaptive Multi-Rate* (GSM-AMR) narrowband (NB) algorithm at 4.75 kbits/s. Importantly, this Adaptive Multi-Rate

(AMR) technology was initially developed for Global System for Mobile communications (GSM) systems, the single most deployed 2G mobile telecommunication standard worldwide and mandatory for 3G cellular systems (Sjoberg et al., 2007). This investigation was an initial study to characterize the differences that exist between clean and coded speech. For this first evaluation, we wanted to make as clear as possible the differences between the two conditions and, therefore, select the lowest bit rate possible.

To examine the effects of compression methods on dysarthric vowel acoustics, we extracted the vowel space area (VSA), using an automated procedure and, then, compared the size and shape across the aforementioned three conditions (original, resampled, and transcoded). The acoustic results were then qualitatively related to listeners' performance on two speech perception tasks: (a) a forced-choice vowel identification task and (b) a phrase transcription task. All listeners were exposed to each condition, from which overall vowel identification accuracy, speech intelligibility, and patterns of lexical boundary errors (LBEs; from the transcription task) were measured. Thus, this study addressed the following research questions: (a) Is VSA differentially affected by the resampling and compression conditions? and (b) Do processing methods affect perceptual outcomes of dysarthric speech, as indexed by vowel identification accuracy, intelligibility, and lexical segmentation patterns?

Method

Speakers and Stimuli

All speech samples were previously recorded and edited for use in a larger study conducted in the Motor Speech Disorders Lab at Arizona State University. Each speaker read words and phrases from visual prompts presented on a computer screen while seated in a sound-attenuating booth, and recordings were made using a high-quality microphone and preamplification, 44.1-kHz sampling rate, and 16-bit pulse code modulation encoding. Recordings were saved directly in a disc for subsequent editing using commercially available software (SoundForge, Sony, 2004) to remove any noise or extraneous speech or silence before or after target utterances. Participants were instructed to read the phrases and words in a "normal, conversational voice." For the purposes of this case study, two speakers with dysarthria were selected. Speakers were male, with chronic dysarthria secondary to traumatic brain injury. Two certified SLPs characterized the speakers' speech patterns. The speech of Speaker 1, a 26-year-old man, was dominated by irregular articulatory breakdowns, imprecise articulation, irregular breath groups, and inappropriate pausing, consistent with a severe ataxic dysarthria. The speech of Speaker 2, a 37-year-old man, was dominated by a strained/strangled vocal quality and a slowed rate, consistent with a moderate spastic dysarthria. The use of these speakers allowed us to evaluate the acoustic and perceptual effects of compression on two speakers with dysarthria, with different predominant speech characteristics.

We recorded 11 vowel tokens embedded in a /bVt/ context (i.e., beat, bat, bit, boot, bought, bet, but, boat, bait, bart, and bert) and a set of phrases constructed for LBE analysis (Liss et al., 2009). Briefly, these phrases contained six syllables and were composed of three to five monosyllabic or disyllabic words, with low semantic transitional probability. The phrases alternated between strong and weak syllables, where strong syllables were defined as those carrying lexical stress. For example, “bolder ground from justice” alternates strong and weak syllables, and “embark or take her sheet” alternates weak and strong syllables. This allows for an examination of lexical segmentation, vowel identification, and overall accuracy, without the influence of contextual information on understanding.

These recordings comprised the original condition, which was then used to create the resampled and trans-coded conditions. The original, high-quality recording had a 44.1-kHz sampling rate and was amplitude normalized. The resampled file had a sampling frequency of 8 kHz. Finally, the transcoded file was processed with a GSM-AMR NB encoder/decoder (REL-11 Version 11.0.0), with a compression rate of 4.75 kbit/s (the decoded file had a sampling frequency of 8 kHz with 16 bits/sample). Briefly, the GSM-AMR-NB is a linear prediction speech compression algorithm that parameterizes speech on 20-ms frames and is widely used in digital cellular networks. We used the AMR-NB source code provided by the International Telecommunications Union to encode speech in the AMR format; a detailed description of the algorithm is available (European Telecommunications Standards Institute, 2002; Meston, 2003; Sjöberg et al., 2007).

Acoustic Measurement

Each speaker's VSA was calculated via an automated method (Sandoval, Berisha, Utianski, Liss, & Spanias, 2013) to assess the impact of resampling and compression algorithms on the acoustics of the speech signal, as represented by vowel formant frequencies. The automated method of calculating VSA takes into account the entire corpus of vowel productions without regard for vowel identity. This is in contrast to traditional VSA calculations, which use the mean first and second formant frequency coordinates for four corner vowels (Fant, 1973). The result is a characterization of the complete shape of the acoustic area bounded by the first and second formants, as opposed to a triangular or quadrangular shape that traditionally characterizes the working vowel space. The VSA estimate is formed by placing the first and second formants of all produced speech samples on the same plot and calculating the area of the subsequently created shape. Visual inspection was conducted to assess qualitative differences in the shape of the vowel space.

Listeners

Twenty listeners recruited from the undergraduate population at Arizona State University participated in the current study. Listeners were native English speakers, passed

a threshold hearing screening, and self-reported nonimpaired cognitive skills. Participants received \$10 as compensation for their participation.

Data Collection Procedures

All procedures were approved by the Institutional Review Board at Arizona State University. For the experimental task, listeners were seated in front of computers located in quiet cubicles. All computers were equipped with Sennheiser HD 280 closed headphones and were calibrated using a digital sound level meter and a flat plate coupler. Volume was set individually on each computer, to 65 dB (A), and participants did not adjust the volume. The experiment was programmed in Alvin (Hillenbrand & Gayvert, 2005) and consisted of two parts: a forced-choice vowel identification task and a phrase transcription task. The order of completion (vowel identification and transcription) was counterbalanced among the participants.

For the vowel identification task, four instances of the 11 bVt tokens in each of the three conditions for each speaker were played, for a total corpus of 264 trials; these were presented randomly and not blocked by speaker. In each trial, a sound file was created with an on-screen selection of 11 bVt options, presented in both phonetic and orthographic representations. For the transcription task, 40 phrases per speaker per condition were presented in random order. This created a total corpus of 240 transcribed phrases per listener. Completion of each task was self-paced, and participants typed their response without any constraints. Instructions informed the participants that the phrases included real English words that were semantically anomalous. Participants were allowed to replay the phrase or vowel one time.

Analyses

Vowel Identification Analysis

Responses from the vowel identification task were analyzed for accuracy and error patterns, in the form of a confusion matrix. This yielded a tally of errors per vowel and the ability to assess the dimensions along which errors differed from the target (front-back and high-low). Quantitative and qualitative error analyses were completed.

Transcription Analysis

The phrase transcriptions collected from the 20 listeners were analyzed and scored by two trained graduate research associates for (a) the number of words correctly transcribed (intelligibility) and (b) the number, type, and location of LBEs. Possible LBEs are (a) insertion of a word boundary before a strong syllable (IS), (b) insertion of a word boundary before a weak syllable (IW), (c) deletion of a word boundary before a strong syllable (DS), or (d) deletion of a word boundary before a weak syllable (DW). Given that English-speaking listeners rely on the occurrence of strong syllables to identify the onset of words in degraded listening conditions, the most common errors are IS and DW, whereas

IW and DS are less common (Liss, Spitzer, Caviness, Adler, & Edwards, 1998). The phrases transcribed by the listeners were specifically constructed to allow for LBE analysis, by equalizing the opportunity for the occurrence of each LBE type. To quantitatively capture the quality of errors, we calculated the proportion of the expected errors (IS and DW errors) to the total number of errors. This is referred to as the metrical segmentation strategy (MSS) ratio (Cutler & Norris, 1988), for which a ratio greater than 0.5 is interpreted as evidence for a perceptual strategy that relies on the identification of strong syllables for lexical segmentation.

Reliability analysis. Two independent transcript codes were compared for reliability. In the few cases that the coded transcripts did not match, the independent scorers convened and came to a consensus.

Statistical analysis. A multivariate, repeated-measures analysis of variance (ANOVA) was conducted to determine if there were statistically significant differences between speakers (1 and 2) and conditions (original, resampled, and transcoded) or an interaction between the two, for overall intelligibility, total LBEs, and MSS ratio.

Results

Acoustic Measurement: VSA

Figure 1 shows the calculated VSA for each speaker across conditions. The VSAs for Speaker 1 (top graph) show no difference between the original and resampled conditions (.17 and .16 [kHz]², respectively). However, the transcoded condition VSA was smaller than both (.14 [kHz]²) and shifted upward in the dimension of the first formant (F1). The VSAs for Speaker 2 (bottom graph) show a slight shape difference between the original and resampled conditions, with divergence of the shape on the dimension of the F1, despite an unchanged VSA of .14 (kHz)² for both conditions. For Speaker 2, the transcoded condition yielded a larger area (.2 [kHz]²) compared with the other conditions, attributable to lengthening of the F2 (second formant) dimension.

Perceptual Data

Vowel Identification Task

Table 1 contains the results of the forced-choice vowel identification task, collapsed by vowel dimensions (front–back, high–low) to facilitate pattern analysis. Results for individual vowels are included in Supplemental Materials S1 and S2. For Speaker 1, the manipulated conditions yielded roughly similar listener performance—within 3 percentage points—for the front vowels (high and low) relative to the original. However, Speaker 1's back vowels were much more vulnerable to the manipulation conditions, particularly the transcoded condition, which yielded the poorest listener performance. Back-high vowels were frequently misperceived as back-low vowels, and back-low vowels were misperceived as front-low vowels.

Speaker 2's productions yielded minimal differences in the listener error patterns, wherein back vowels (both high and low) and front-low vowels were perceived similarly,

within 3 percentage points, of the original recording. However, interestingly, the manipulation conditions improved listener performance for front-high vowels.

Transcription Task: Words Correct (Intelligibility)

Figure 2 shows the mean intelligibility on the transcription task, for each speaker in each condition. Overall, intelligibility of phrases revealed no change between the original (24%) and resampled (24%) conditions for Speaker 1, with a slight drop in accuracy for the transcoded (20%) condition. Speaker 2, on the other hand, showed a small change in intelligibility from the original (47%) to the resampled (43%) condition, with a dramatic decrease in intelligibility in the transcoded condition (33%). A multivariate, repeated-measures ANOVA revealed a significant effect of speaker, $F(2, 14) = 206.09, p < .001, \eta^2 = .644$, and condition, $F(2, 14) = 19.10, p < .001, \eta^2 = .251$. Follow-up comparisons revealed statistically significant differences between the original and transcoded ($p < .001$) and resampled and transcoded ($p < .001$) conditions, but no difference between the original and resampled conditions ($p = .2119$). In addition, the interaction was significant, $F(2, 14) = 6.97, p < .001, \eta^2 = .109$. There were significant differences for both patients between the original and transcoded conditions and a significant difference between the intelligibility of the two patients in the transcoded condition.

Transcription Task: LBEs

Figure 3 shows the total number of LBEs from the transcription task for each speaker in each listening condition. The total number of LBEs quantifies segmentation accuracy. For Speaker 1, there was a difference of 40 errors between the conditions that have the largest number of errors (resampled) and least number of errors (original). Resampled and transcoded conditions yielded an equivalent number of errors. For Speaker 2, a more interesting pattern appeared where the change in number of LBEs more directly reflected the changes in intelligibility. That is, the condition with the least number of errors had the highest intelligibility (original), and condition with the highest number of errors had the lowest intelligibility (transcoded). The magnitude of the effect was also reflected. A multivariate, repeated-measures ANOVA revealed a significant effect for speaker, $F(2, 14) = 31.85, p < .001, \eta^2 = .216$, a nonsignificant effect for condition, $F(2, 14) = 0.89, p = .42$, and a nonsignificant interaction, $F(2, 14) = 0.17, p = .85$.

Transcription Task: MSS Ratio

Figure 4 shows the MSS ratio for each speaker in each listening condition. MSS ratio quantifies the quality of segmentation accuracy or inaccuracy. An MSS ratio of above 0.5 suggests a stress-based segmentation strategy. On average, the MSS ratios suggest that neither condition nor speaker influenced the use of stress cues to segment the speech stream. A multivariate, repeated-measures ANOVA revealed a nonsignificant effect for both speaker, $F(2, 14) = 1.71, p = .19$, and condition, $F(2, 14) = 0.64, p = .53$, and a nonsignificant interaction, $F(2, 14) = 0.63, p = .54$.

Figure 1. The vowel space area (VSA in $[\text{kHz}]^2$) for Speakers 1 (top graph) and 2 (bottom graph) in each condition. The solid green represents the VSA associated with the original. The dashed blue refers to the resampled condition, and the dotted red refers to the transcoded condition. The numerical VSA (in $[\text{kHz}]^2$) is listed for each condition, for each speaker. The axes are oriented such that the plots have the same orientation as the standard International Phonetic Alphabet vowel trapezium (Sandoval et al., 2013). There are different scales for the dimension of the second formant to allow for demonstration of individual speaker changes.

Table 1. Confusion matrices for front-back confusion for each condition: (a) original, (b) resampled, and (c) transcoded for each speaker (top: Speaker [Sp.] 1; bottom: Speaker 2); number indicates percentage of tokens identified within each class (front-high [F-H]: /i/, /ɪ/, and /e/; front-low [F-L]: /E/ and /ae/; back-high [B-H]: /u/ and /o/; back-low [B-L]: /ʌ/ and /a/) out of the corpus of 80 classifications per vowel per speaker.

Sp. 1	Original (perceived)				Resampled (perceived)				Transcoded (perceived)			
Actual	F-H	F-L	B-H	B-L	F-H	F-L	B-H	B-L	F-H	F-L	B-H	B-L
F-H	90.8		9.2		93.8	5.4	0.8		90	8.3	1.7	
F-L		95.6	4.4		3.2	96.2	0.6		1.3	98.1	0.6	
B-H	3.3	1.3	91.3	4	2.7	2	90.6	4.7	3.3	2	80.7	14
B-L	0.6	19.1		80.3	0.7	24.5		74.8		25.2	1.3	73.5
Sp. 2	Original (perceived)				Resampled (perceived)				Transcoded (perceived)			
Actual	F-H	F-L	B-H	B-L	F-H	F-L	B-H	B-L	F-H	F-L	B-H	B-L
F-H	78.6	21		0.4	85.7	12.6	0.4	1.3	86.8	10.7	1.7	0.8
F-L	0.7	97.4		1.9	1.2	96.3		2.5		95.5		4.5
B-H			100		0.6		99.4		0.6		99.4	
B-L		4.6		95.4		1.3		98.7		4	1.4	94.6

Note. Confusions in “er” or “ar” were omitted. Accurate identification is noted in bolded text.

Discussion

The current study builds upon past research that suggests that there is a complex interplay between speech compression algorithms on nonhealthy speech (Parsa & Jamieson, 2000), accented speech (Itani & Paulikas, 2008), and even different languages (Alencar & Alcaim, 2009; Paulikas & Itani, 2009). The results of the current study are not incompatible with a previous report of the effects of speech coding algorithms on understanding nonsense words produced by individuals with disordered speech (Jamieson et al., 2002). In the same report, it was revealed that intelligibility for vowels was not directly related to consonants, with vowels being more easily identified; this is consistent with our findings that vowel identification was much higher than the intelligibility of connected speech. The current report has expanded on these past reports with further characterization of the speech patterns (i.e., descriptions of dysarthria), multiple outcomes of intelligibility, and, most importantly, by utilizing connected speech.

This case study explores two speakers diagnosed with dysarthria: one who was severely unintelligible and presented with features consistent with ataxic dysarthria and a large VSA (Speaker 1) and a speaker who was moderately unintelligible, with features consistent with spastic dysarthria and a relatively smaller VSA (Speaker 2). Due to the non-uniform acoustic and perceptual findings revealed for the two speakers with dysarthria, we offer insight into the posed research questions and then discuss the results for each speaker in greater detail.

Given that there were different effects for each speaker, it appears that VSA is differentially affected by the resampling and compression conditions. This reflects the answer to the first posed research question. For Speaker 1, there were no differences in the VSA associated with resampling compared with the original; however, transcoding compressed the vowel space. For Speaker 2, resampling also had no effect, whereas transcoding increased the calculated VSA.

Our second research question was whether processing methods affect perceptual outcomes of dysarthric speech, as indexed by vowel identification accuracy, intelligibility,

Figure 2. Mean intelligibility (% words correct) on the transcription task, with standard error, for each speaker in each listening condition.

Figure 3. Total number of lexical boundary errors (LBEs) from the transcription task for each speaker in each listening condition.

Figure 4. MSS ratio, derived from lexical boundary errors of all listeners in the transcription task, with standard error, for each speaker in each listening condition. MSS = metrical segmentation strategy.

and lexical segmentation patterns. Overall, the results suggest that transcoding may affect identification of vowels, intelligibility, and the quantity, but not quality, of segmentation errors.

Speaker 1

The changes seen in the VSA for Speaker 1 parallel the lack of change in overall intelligibility (resampled) and slight decrease in intelligibility (transcoded). Further research will assess the threshold at which a change in VSA yields perceptual consequences and the relationship of the magnitude of change in the vowel space (.17 in original to .14 in transcoded for Speaker 1) to that in intelligibility (24% in original to 20% in transcoded for Speaker 1). The vowel identification data, specifically the error patterns, further reflect the change to the acoustic vowel space. Consistent with the upward shift of F1 seen in the VSA in transcoding, there was reduced vowel identification accuracy of back-high vowels (~10%), with the majority of the perceptual errors occurring along the F1 (high–low) dimension. There were no noted effects in LBEs or segmentation strategy associated with either resampling or transcoding.

Speaker 2

For Speaker 2, transcoding expanded the vowel space (see Figure 1, bottom graph); specifically, the GSM-AMR algorithm expanded the VSA along the F2 (second formant) dimension, which is the acoustic analogue to tongue excursion (i.e., front–back). This expansion in vowel space corresponded to improved vowel identification accuracy, particularly for front-high vowels. Interestingly, the GSM-AMR algorithm reduced intelligibility significantly (transcoded). In addition, listeners made the largest amount of LBEs (see Figure 3) in the transcoded condition. Although transcoding resulted in an expanded vowel space, which may have improved vowel perception, it possibly negatively affected the fidelity of the consonantal information. Reduced consonant identification is a possible explanation for

the increased segmentation errors and reduced phrasal intelligibility (Ansel & Kent, 1992); systematically investigating this will be the focus of future research.

General Discussion

Importantly, coding with the GSM-AMR-NB and other codecs reconstructs the speech signal using a model trained on healthy speech. The algorithm contains multiple codebooks for efficiently compressing parameters of the input speech. These codebooks are composed of entries learned from healthy speech corpora. It was conjectured that, for dysarthric speech inputs, the codec remaps formants of the disordered speech signal to the closest companions in the codebook from healthy speech. This resulted in a modified representation of the speech signal, as represented by the altered VSA. As demonstrated by the differential effects on the VSAs of Speakers 1 and 2, the effects can either be helpful or harmful (i.e., the “closest” healthy match in the reconstruction is farther away from the original speech sample and results in the selection of an alternate phoneme).

Resampling, a necessary first step in these speech codecs, resulted in an overall decrease in intelligibility (4% points) for Speaker 2. Perceptually, the consequences of resampling lead to poor resolution of consonants, which was the presumed reason for the overall decreased intelligibility. This was likely secondary to a loss of high-frequency information from the low-pass filtering inherent in resampling. The codec exacerbates these differences, as it introduces artifact between 3.4 and 4 kHz and subsequent coding distortion. This was supported by an additional 10% decrement in intelligibility for Speaker 2.

Limitations

There are, of course, limitations to this study. Inherent in a case series is a small sample size that is not representative of the range of patients seen in clinical practice. It is possible that floor effects are present given the original severity of Patient 1 and that patients with mild–moderate dysarthria would be differentially affected. Moving forward, it will be important to assess these findings in a wide range of dysarthria types and severities. We also used an automated VSA that captures the entirety of the vowel space (Sandoval et al., 2013). Although this has not extensively been tested on dysarthric speech, it is a robust method that uses iterative algorithms; any unintended consequences from the automation process should be removed by such an iterative process. Finally, we utilized nonexperienced, undergraduate listeners for this experiment; however, it is unlikely that having naive listeners impacted the results of the study. Past research shows that level of clinical experience may not directly influence listener perception (Lansford, Berisha, & Utianski, 2016). Nonetheless, it will be important to replicate these findings with experienced speech-language pathologists, who will likely be the individuals providing services via telepractice.

Conclusions and Future Directions

In light of the results of this case study, as well as the growing practice of telemedicine, it is critical to be mindful of apparent changes in intelligibility secondary to compression algorithms and resampling utilized in telepractice. It certainly warrants caution to ensure changes documented in patients are secondary to improvements or progression of the communication disorder and not a byproduct of transmission artifact. Here, we selected the lowest bit rate possible (4.75 kbits/s), to isolate the effects of the compression algorithm. Future studies should analyze the error patterns for the different bit rates.

Overall, the current study suggests that changes in intelligibility may be an artifact of the compression algorithms utilized in transmitting speech from client to practitioner and may be affected by the pattern of speech impairment. There seems to be a complex interplay between the dysarthria and the compression algorithm. It is possible that different disruptions to rhythm (for instance, even/equal stress patterns, monopitch, monoloud) are differentially affected. In the future, research will work toward characterizing the acoustic parameters that relate to the changes in intelligibility. That is, what elements of the speech signal, beyond VSA, change along with improved or diminished perception? It will also be critical to systematically investigate whether reduced intelligibility was indeed due to missed consonantal information rather than vowels. This, and other future works with a larger cohort of dysarthric speakers, will be critical for guiding the development of algorithms for use in telemedicine.

Acknowledgments

This research was supported in part by National Institute on Deafness and Other Communication Disorders Grants 2R01DC006859 (Liss) and 1R21DC012558 (Liss and Berisha). Rene L. Utianski, Julie M. Liss, and Visar Berisha received research support from the National Institutes of Health. Portions of this work were presented at the 164th meeting of the Acoustical Society of America and were awarded second place in the Best Paper Award competitions for Students and Young Presenters in Speech Communication.

References

- Alencar, V. F. S. D., & Alcaim, A. (2009, August). *On the performance of ITU-T G.723.1 and AMR-NB codecs for large vocabulary distributed speech recognition in Brazilian Portuguese*. Paper presented at the 2009 Second International Conference on the Applications of Digital Information and Web Technologies, London, United Kingdom.
- American Speech-Language-Hearing Association. (2012). Telepractice: Research trends. *The ASHA Leader*, 17(12), 24–25.
- Ansel, B. M., & Kent, R. D. (1992). Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy. *Journal of Speech and Hearing Research*, 35(2), 296–308.
- Barreto, A. N., & Pirhonen, R. (2006, May). *Capacity increase in GSM networks using source-adaptive AMR*. Paper presented at the 2006 IEEE 63rd Vehicular Technology Conference, Melbourne, Australia.
- Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of experimental psychology: Human perception and performance*, 14(1), 113–121.
- European Telecommunications Standards Institute. (2002). *Universal Mobile Telecommunications System (UMTS): AMR speech codec: Transcoding functions (ETSI TS 126 090 V5.0.0)*. Sophia Antipolis, France: European Telecommunications Standards Institute.
- Fant, G. (1973). *Speech sounds and features*. Cambridge, United Kingdom: MIT Press.
- Guillemin, B. J., & Watson, C. (2008). Impact of the GSM mobile phone network on the speech signal: Some preliminary findings. *International Journal of Speech, Language and the Law*, 15(2), 198–218.
- Hillenbrand, J. M., & Gayvert, R. T. (2005). Open source software for experiment design and control. *Journal of Speech, Language, and Hearing Research*, 48(1), 45–60.
- Holma, H., Melero, J., Vainio, J., Halonen, T., & Makinen, J. (2003, October). *Performance of adaptive multirate (AMR) voice in GSM and WCDMA*. Paper presented at the 57th Vehicular Technology Conference, Jeju, South Korea, South Korea.
- Itani, M., & Paulikas, S. (2008). Influence of languages on CELP codecs performance. *Information Technology and Control*, 37(2), 141–144.
- Jamieson, D. G., Parsa, V., Price, M. C., & Till, J. (2002). Interaction of speech coders and atypical speech effects on speech intelligibility. *Journal of Speech, Language, and Hearing Research*, 45(3), 482–493.
- Lansford, K. L., Berisha, V., & Utianski, R. L. (2016). Modeling listener perception of speaker similarity in dysarthria. *The Journal of the Acoustical Society of America*, 139(6), EL209–EL215.
- Liss, J. M., Spitzer, S., Caviness, J. N., Adler, C., & Edwards, B. (1998). Syllabic strength and lexical boundary decisions in the perception of hypokinetic dysarthric speech. *The Journal of the Acoustical Society of America*, 104(4), 2457–2566.
- Liss, J. M., White, L., Mattys, S. L., Lansford, K., Lotto, A. J., Spitzer, S., & Caviness, J. N. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. *Journal of Speech, Language, and Hearing Research*, 52(5), 1334–1352.
- Meston, R. (2003, July). Sorting through GSM codecs: A tutorial. *EE Times: Connecting the Global Electronics Community*. Retrieved from https://www.eetimes.com/document.asp?doc_id=1202660
- Parsa, V., & Jamieson, D. G. (2000). Identification of pathological voices using glottal noise measures. *Journal of Speech, Language, and Hearing Research*, 43(2), 469–485.
- Paulikas, S., & Itani, M. (2009, December). *Assessment of CELP codecs quality in multi-lingual environment*. Paper presented at the 2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Ajman, United Arab Emirates.
- Sandoval, S., Berisha, V., Utianski, R. L., Liss, J. M., & Spanias, A. (2013). Automatic assessment of vowel space area. *The Journal of the Acoustical Society of America*, 134(5), EL477–EL483.
- Sauter, M. (2010). *From GSM to LTE: An introduction to mobile networks and mobile broadband*. West Sussex, United Kingdom: Wiley.
- Sjoberg, J., Westerlund, M., Ericsson, L., Lankaniemi, A., Nokia, Xie, Q., & Motorola. (2007). *RTP payload format and file storage format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) audio codecs*. Retrieved from <https://www.rfc-editor.org/info/rfc4867>

Spitzer, S., Liss, J. M., & Mattys, S. L. (2007). Acoustic cues to lexical segmentation: A study of resynthesized speech. *The Journal of the Acoustical Society of America*, 122(6), 3678–3687.

Thorpe, L. A. (1999). Subjective evaluation of speech compression codecs and other non-linear voice-path devices for telephony applications. *International Journal of Speech Technology*, 2(4), 273–288.