USING SPEECH TRANSFORMATION TO INCREASE SPEECH INTELLIGIBILITY FOR THE HEARING- AND SPEAKING-IMPAIRED

Alexander Kain and Jan van Santen

Center for Spoken Language Understanding
Oregon Health & Science University
Portland, Oregon 97239, USA

BioSpeech, Inc.
940 Upper Devon Lane
Lake Oswego, Oregon 97034, USA

ABSTRACT

We present two speech transformation approaches designed to increase the intelligibility of speech. The first approach is used in the context of increasing the intelligibility of conversationally spoken speech for hearing-impaired listeners. An initial experiment showed that a relatively simple mapping function can map spectral features of conversationally spoken speech closer to context-equivalent spectral features of clearly spoken speech. The second approach aims to increase the intelligibility of speaking-impaired individuals by the general population. Results of listening tests indicated that although an intelligibility increase was not achieved, listeners preferred the transformed speech of the proposed system over that of an alternative system.

***Index Terms*—** speech modification, hearing aid, speaking aid.

1. INTRODUCTION

A speech transformation system has speech as both its input and its output (as opposed to speech recognition or text-to-speech synthesis). Systems have focused on altering speaker characteristics (changing the perceived speaker), voice characteristics (changing the voice style of the same speaker), emotions, and other aspects. We report here on experiments involving two speech transformation approaches that aim to increase the intelligibility of their input speech. The first approach has as its input conversational speech of any speaker of the general population, and presents its enhanced output to a hearing-impaired listener. The second approach transforms speech of an impaired speaker into an enhanced version for presentation to the general population. The first transformation approach has potential applications as an assistive hearing-aid device, while the second has potential to play the role of an assistive speaking-aid.

2. IMPROVING THE INTELLIGIBILITY OF CONVERSATIONAL SPEECH

Approximately 28 million people in the United States have some degree of hearing loss, with 40–45% of the population over 65, and about 83% of those over 70, classified as hearing impaired. Elderly listeners often have an especially difficult time understanding speech in noise or under distracting conditions. Wearing a hearing aid is one of the most often-used strategies that can partially compensate for a hearing impairment. The primary benefit of hearing aids is to restore hearing loss resulting from reduced sensitivity, by amplifying signal energy in one or more frequency bands. No attempts are made to perform prosodic or fine-grained spectral modifications, even though it is known that increased speech intelligibility can be obtained by processes distinct from simply regulating the energy of the speech signal. For example, speakers naturally adopt a special speaking style when aiming to be understood by listeners who are moderately impaired in their ability to understand speech due to hearing loss, the presence of background noise, or both. This style has been termed "clear" (CLR) [e. g. 1]. In contrast, speech intended for a normal-hearing listener in a quiet environment is commonly referred to as "conversational" (CNV). The intelligibility of CLR speech is higher than that of CNV speech, as measured in listeners of different age groups, with normal and impaired hearing abilities, using different types of speech materials, and in environments with different types of background noise [for a brief overview, see 2].

Previous research has examined acoustic differences between CNV and CLR speech. The following prosodic features have been found to distinguish CLR speech from CNV speech: (1) the fundamental frequency (F0) is typically increased in range and mean, (2) the consonant-vowel energy ratio (CVR) is increased, particularly for stops and/or affricates, (3) phoneme durations are prolonged, especially in the tense vowels /i/, /u/, /A/, and /O/, (4) pauses are longer in duration and occurred more frequently, and (5) the speaking rate is significantly decreased. The following spectral features distinguish CLR speech from CNV speech: (1) vowel formant frequencies show expanded vowel spaces for lax vowels, (2) long-term spectra have increased energies at higher frequencies (1000–3150 Hz), (3) alveolar flaps occur less often and consonant stops tend to be released with following aspiration, and (4) some speakers exhibit increased modulation indices for low modulation frequencies up to 4 Hz.

Current hearing aid systems focus on amplifying the speech signal in one or more frequency bands. However, this approach does not address important problems that may be encountered by users, especially elderly listeners, who have more difficulty than younger listeners in understanding rapid speech due to decreased auditory processing capabilities or reduced working memory capacity. Motivated by these findings, researchers have developed signal-processing algorithms to increase the intelligibility of speech independent of amplification. Modifications included decreasing the rate of speech by inserting pauses, modifying phoneme durations, and enhancing the consonant-to-vowel energy ratios. Of these, only one study showed

The research described in Section 3 was funded by U. S. National Institute of Health STTR grant 1R41DC007240-01A2. Oregon Health & Science University, A. Kain, and J. van Santen have a significant financial interest in BioSpeech, Inc., a company that may have a commercial interest in the results of this research and technology. This potential conflict was reviewed and a management plan approved by the Conflict of Interest in Research Committee and the Integrity Program Oversight Council was implemented.

a statistically significant increase in intelligibility at the sentence level, by amplifying the energy of specific consonants. Ultimately, the causal relationship between sets of acoustic features and speech intelligibility is not yet known.

To address this, recent studies reported on experiments that measured the degree of contribution of six high-level acoustic features to intelligibility, by applying certain CLR features to CNV speech [2, 3]. This has been accomplished by using a "hybridization" algorithm that (1) extracts CNV and CLR features from the same sentences spoken in both CNV and CLR styles, then (2) constitutes a "hybrid" (HYB) feature set from a particular subset of CLR features and from the complementary subset of CNV features, and finally (3) synthesizes HYB sentences from the HYB features. Testing the intelligibility of the HYB speech in noise indicated that the two main sources of increased intelligibility of CLR speech for one particular speaker are the spectrum and phoneme duration, and not the pausing patterns, F0, energy, or phoneme sequence. For example, in one experiment, the intelligibility of CNV was increased from 72% to 82%, using CLR spectrum and phoneme duration features. The hybridization algorithm is equivalent to transforming CNV speech with an "oracle" mapping function, thus simulating maximum performance levels of an automatic transformation system. In the remainder of this section, we propose and test the feasibility of a spectral mapping component of such a speech transformation system.

2.1. Speech Corpus

The text material consisted of 70 phonetically-balanced sentences from the set of IEEE Harvard Psychoacoustic Sentences [4]. One male, a native speaker of American English, was recruited as a speaker. We first recorded the sentences spoken in the CNV speaking style, then the same sentences spoken in the CLR speaking style. When recording CNV speech, the speaker was instructed to speak in the way that he communicates in his daily life. When recording CLR speech, he was instructed to speak clearly, as he would when communicating with a hearing-impaired listener. A technician listened to each sentence, and the speaker was asked to record a sentence again when pronunciation or style were not satisfactory. Initial estimates of phoneme identities (including non-speech events) and boundaries in each waveform were obtained using an existing forced-alignment system [5]. Then a trained labeler checked and adjusted phoneme identities and boundaries manually.

2.2. Spectral Mapping Experiment

We tested the hypothesis that a relatively simple transform exists that maps CNV spectral features closer toward CLR spectral features (motivated by earlier encouraging hybridization results [2]). First, CNV and CLR speech sentences were aligned phonetically, since the speaker often pronounced the same sentences differently, depending on the speaking style (e. g. CLR speech had more pauses and unvoiced plosive releases). To accomplish the alignment, a phoneme feature table was created specifying voicing, manner, place, and height features, with one 4-dimensional vector for each phoneme. Each phonetic symbol in both label sequences was assigned its associated feature vector, resulting in two feature matrices. Then, dynamic time warping was used to find an optimal alignment path between the two matrices that resulted in the minimum Euclidean distance between the corresponding phonetic features. As a result, each phoneme in one speaking style was associated with one phoneme in the other speaking style either by a perfect match or a best-fit match. In those cases where a one-to-one mapping was not possible, the phoneme was considered to be an insertion/deletion.

We extracted 20^{th}-order Linear Prediction Coefficients (LPC) from asynchronous, 50% overlapping, Hanning-windowed, 25 msec frames of the speech waveforms (sampled at 16 kHz) in the speech corpus, and then converted the coefficients into Line Spectral Frequencies (LSF). Iterating over all sentences, we accumulated context-equivalent CNV-LSF and CLR-LSF vectors for (1) vowels, or (2) all phonemes that had a one-to-one mapping, stretching the shorter sequence to match the longer sequence when phoneme durations differed.

We designated 49 sentences of the corpus (70%) for training, and the remaining sentences for testing. We estimated a Gaussian mixture regression model [6] on the joint density of the CNV-LSF and CLR-LSF vectors (3174 and 8653 vectors for vowels and all phonemes, respectively) of the training data. Evaluating on the test set, the log-spectral distance between mapped CNV spectra and associated CLR spectra of vowels was 4.56 dB, as compared to the original distance of 5.36 dB between unmapped CNV spectra and associated CLR spectra, and 4.57 dB versus 5.17 dB for all phonemes. The optimal number of mixture components was between two and four. Thus, the mapping function was able to produce mapped CNV spectral features that moved closer toward CLR spectral features. We transformed a number of example sentences from the test set, using a LPC vocoder synthesis scheme and the spectral mapping function trained on all phonemes; informal listening tests indicated that the transformed speech had very high quality. Further experiments involving different speech representations and formal intelligibility tests are planned.

3. IMPROVING THE INTELLIGIBILITY OF DYSARTHRIC SPEECH

Dysarthria is a speech motor disorder usually resulting in a substantive decrease in speech intelligibility by the general population. In this section, we discuss a speaking-aid system with the aim of improving the intelligibility of talkers with dysarthria. In overview, the system first enrolls a new source dysarthric speaker by recording his or her speech while reading of a word list. The system then analyzes these recordings and computes mapping function parameters. During normal use, the dysarthric speaker talks with a microphone in place, and when he or she is finished the system plays the transformed speech over amplified speakers. We refer to this mode of operation as "interpreter mode" because of the similarities to foreign language interpretation.

The key idea of our approach is to improve intelligibility by analysis, mapping, and synthesis of a small set of perceptually-relevant speech features. The mapping step consists of moving the dysarthric features towards known good target features by means of a trained mapping function. The particular choice of speech features is motivated by the need to represent speech intelligibly, but not necessarily very naturally or with the dysarthric speaker's own voice. At the same time, the number of training parameters should be kept small to allow training of the transformation function with a relatively small amount of training data. For these reasons, the speech features and synthesis methods used are similar to a formant speech-synthesis approach, producing highly intelligible and controllable speech from a compact representation. The system consists of three major parts: a speech analyzer (used both during training and transformation), a feature mapping learner (training), and a speech transformer. Enrollment consists of analysis and training, whereas transformation consists of speech analysis, feature mapping, and speech synthesis.

Figure 1. BioSpeech system components: Headset (left) is attached via USB adapter to PC (middle, belt attachment not shown), which is attached to amplified speaker (bottom, with adjustable belt).

In a recent study, the intelligibility of dysarthric vowels of one speaker was improved from 48% to 54% [7]. Improvement was obtained by transforming monophthong vowels of a speaker with dysarthria to more closely match the vowel space of a non-dysarthric (target) speaker, using a feature set consisting of vowel duration and formant F1, F2, and F3 stable points [see also 8]. We describe here extensions to this system, with the aims of transforming any vowel (including diphthongs) and distinguishing vowels from consonants automatically. It was also required to implement the system to run on a wearable platform (see Figure 1), consisting of a very small personal computer (OQO model 02), a lightweight headset (Sennheiser ME3), an amplified speaker (ChatterVox). A final goal was to evaluate system performance extensively, using both listening experiments and product surveys, comparing it to a commercially available system, the SpeechEnhancer [9].

3.1. Speech Corpus

As text material, we created a 336 word list, consisting of consonant (C) to vowel (V), VC, and CVC words. We recorded speech waveforms of 5 dysarthric (ataxic, flaccid, hyperkinetic, mixed types; average intelligibility was 58%) and 2 normal speakers uttering the text material using a special purpose graphical user interface we developed for this task. After the recording process, we used a force-alignment system [5] to create initial phoneme boundaries (phoneme identities were given by the text material itself), which were subsequently checked and adjusted manually by an experienced labeler for maximum accuracy. About two thirds of the data was used for training, and about one third for testing.

3.2. Vowel Region Localization

The system has to automatically locate the vowel region in the speech input; for this purpose, we trained a broad-category recognizer for each dysarthric speaker, using broad classes {/j/ /w/ /l/ /9r/ B C V clo}, where B = burst, C = consonant which does not include {/j/ /w/ /l/ /9r/ B}, V = vowel, and clo = pause or burst closure. In testing we mapped B to C, for a total of 7 output categories. (We trained B as distinct from C because the acoustics are different.) Training was done using up to 4000 examples of each class. The features were standard Mel-frequency warped cepstral coefficient features using 22 filters and a window size of 16 msec. The classifier used context-dependent categories, for a total of 38 output classes. The phoneme labels used in training were obtained from the forced-alignment performed earlier. Recognition of the test data used the following grammar: {C1 V | V C2 | C1 V C2 | V c B | C1 V c B} where C1 is any consonant, including a burst, C2 is any consonant other than {B /j/ /w/}, and c is a burst closure. After recognition, C1, C2, and cB were all mapped to C. Testing the vowel region localizer performance it was found that the average absolute distance between the predicted vowel boundaries and the actual boundaries was 22 msec, as compared to 50 msec using a previous approach involving isotonic regression [7].

3.3. Enrollment and Transformation

Enrollment consisted of two steps: feature analysis and subsequent training of a mapping function. During the first step, we aligned the enrollment (source) speaker's speech with a target speaker of the same gender and extracted formant frequencies frame-by-frame from both source and target speakers. We then fit a joint-density Gaussian mixture regression model [6] on the training data.

During transformation, an entire utterance was analyzed to obtain formant frequency values, to be used as input to the formant modification operation. Unvoiced frames of speech, however, were passed directly to the output. Additionally, any speech frequencies above 4 kHz were passed through unmodified, to the output signal. To generate the voiced regions of the transformed speech, the system performed modifications to the energy and formant features (estimated using ESPS algorithms), and generated a new F0 trajectory from the CVC boundary information. Energy modification was applied because the dysarthric speech often contained significant energy flutter (variations in energy), likely caused by high levels of "vocal fry". Similarly, the F0 trajectory of a dysarthric speaker often contained significant jitter (variations in F0). We discarded the original F0 values, which were often estimated with large errors, in favor of a synthetic F0 contour, generated by a simple superpositional intonation model. Formants were modified by estimating formant vectors frame-by-frame for the entire dysarthric vowel, mapping all vectors to the transformed formant vectors using the trained joint-density Gaussian mixture regression model, and then creating the transformed formant trajectory. This allowed us to transform monophthong and diphthong vowels alike; the previous approach was limited to only monophthong vowels.

The transformation system operated in approximately $0.2\times$ real-time; in other words, an utterance that was approximately 2 seconds long (a very slowly spoken single syllable word) will be processed in 400 msec (pipelining the system would further reduce this delay).

3.4. Listening Tests

Fourteen listeners participated in our listening tests, all native speakers of American English and with self-reported normal hearing.

	Original	Original+ BioSpeech	Original+ SpeechEnh.	Original+ Original
AAG	82	76	79	80
JMJ	68	68	64	69
Total	75	72	71	74

Table 1. Intelligibility scores in percent.

Ages ranged from 29–63, with an average age of 43. They listened to the stimuli using high-quality speakers in a quiet office environment. The BioSpeech and SpeechEnhancer system outputs were previously recorded from their respective portable speakers using a high-quality microphone. Tests were administered via a computerized experimental setup controlled by mouse.

In a preference test, listeners made a graded forced choice on a five-point scale (where -2 and $+2$ indicated strong preferences, -1 and $+1$ weaker preferences, and 0 indifference) between two speech samples. One of the stimuli was transformed using the BioSpeech system and the other by the SpeechEnhancer. A male voice and a female voice were selected (voices AAG and JMJ, respectively). Result indicated an overall preference for the BioSpeech system over the SpeechEnhancer. The raw scores were processed such that positive values indicate a preference for the BioSpeech system and negative values a preference for the SpeechEnhancer. An average value of 0.433 ($t_{13} = 2.15, p < 0.05$) was obtained. However, these results were strongly dependent on speaker, with an average value of 0.689 ($t_{13} = 3.06, p < 0.01$) for the male speaker and 0.183 for the female speaker ($t_{13} = 0.9$, not significant). This difference may depend on a gender difference or on the male speaker having substantially lower intelligibility (50%) than the female speaker (80%). (Note, however, that the intelligibility test indicated that the male speaker was more, and not less, intelligible than the female speaker; this discrepancy may be due to the clinical intelligibility assessing both consonants and vowels.)

In an intelligibility test, stimuli were presented in one of the following conditions: (1) the original dysarthric speech, simulating the absence of any enhancement, (2) the original speech followed by the BioSpeech system output, (3) the original speech followed by the SpeechEnhancer system output, and (4) the original followed by the original speech again, to measure the effect of mere repetition. Listeners were asked to indicate which of four vowels they heard, for all conditions. Results indicated no significant differences between the BioSpeech and SpeechEnhancer products, but regrettably also not between the systems and the original speech. In summary, neither system enhanced intelligibility for these two speakers, while the two products performed almost equally to each other.

3.5. Survey

Two videos were produced of individuals interacting with simulated versions of the BioSpeech system. Ten adults viewed the videos, inspected the hardware, and completed a survey developed based on assistive technology outcomes efforts [10]. Seven participants had amyotrophic lateral sclerosis and were moderately to severely speech impaired; three participants were family members and primary communication partners.

The participants answered questions relating to the following three survey topics: design and appearance, function, and personal preference. In summary, the survey indicated that adults with moderate to severe dysarthria think that the proposed system is a viable, reliable, attractive and creative option to augment speech.

4. CONCLUSION

Speech transformation systems have tremendous potential to contribute significantly to the quality of life of individuals with communication disorders. In this paper, we reported on research with the goal of increasing intelligibility for impaired listeners or impaired speakers. A first approach aims to transform conversational speech of any speaker to more closely resemble clearly spoken speech, with the potential of being used in assistive listening devices. Initial experiments showed that a relatively simple mapping function was able to map spectral features of conversationally spoken speech closer to those of clearly spoken speech.

A second approach aimed at improving the intelligibility of speakers with dysarthria by the general population, but did not succed in demonstrating a significant improvement for the speakers used in our evaluation. This may be because the transformation system is optimal when the vowel space is strongly deviant, in combination with relatively low formant variability. However, listeners preferred the speech as modified by the proposed system as compared to a competing system; presumably, this is due to the elimination of hoarseness and other potentially unattractive features of the original speech in the output speech.

5. REFERENCES

[1] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *Journal of Speech and Hearing Research*, vol. 29, pp. 434–446, 1986.

[2] A. Kain, A. Amano-Kusumoto, and J.-P. Hosom, "Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility," *Journal of the Acoustical Society of America*, vol. 124, no. 4, October 2008.

[3] A. Kusumoto, A. Kain, J.-P. Hosom, and J. van Santen, "Hybridizing conversational and clear speech," in *Proceedings of Interspeech*, August 2007.

[4] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silberger, G. E. Urbanek, and M. Weinstock, "IEEE Recommended practice for speech quality measurements," *IEEE Trans. on Audio Electroacoustics*, vol. 17, pp. 227–246, 1969.

[5] J.-P. Hosom, "Automatic phoneme alignment based on acoustic-phonetic modeling," in *Proceedings of ICSLP*, Boulder, CO, 2002, vol. 1, pp. 357–360.

[6] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of ICASSP*, May 1998, vol. 1, pp. 285–288.

[7] A. Kain, J.-P. Hosom, X. Niu, J. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 743–759, September 2007.

[8] A. Kain, X. Niu, J.-P. Hosom, Q. Miao, and J. van Santen, "Formant re-synthesis of dysarthric speech," in *Proceedings of the 5th ISCA Workshop on Speech Synthesis*, June 2004.

[9] Voicewave Technology Inc., "The Speech Enhancer," `www.speechenhancer.com`, September 2008.

[10] M. Scherer, *Living in the State of Stuck: How Assistive Technology Impacts the Lives of People with Disabilities*, Brookline Books, Cambridge, MA, 2000.