Hybridizing Conversational and Clear Speech

Akiko Kusumoto, Alexander B. Kain, John-Paul Hosom, Jan P. H. van Santen

Center for Spoken Language Understanding
OGI School of Science & Engineering
Oregon Health & Science University
20000 NW Walker Road, Beaverton, OR 97006, USA

Abstract

"Clear" (CLR) speech is a speaking style that speakers adopt to be understood correctly in a difficult communication environment. Studies have shown that CLR speech, as opposed to "conversational" (CNV) speech, has significantly higher intelligibility in various conditions. While many differences in acoustic features have been identified, it is not known which individual feature or combinations of features cause the higher intelligibility of CLR speech. The objectives of the current study are to examine whether it is possible to improve speech intelligibility by approximating CLR speech features and to determine which acoustic features contribute to intelligibility. Our approach creates speech samples that combine acoustic features of CNV and CLR speech, using a hybridization algorithm. Results with normal-hearing listeners showed significant sentence-level intelligibility improvements of 11–23% over CNV speech when replacing certain acoustic features with those from CLR speech.

Index Terms: Speech processing, Speech intelligibility, Speech communication, Hearing aids.

1. Introduction

Approximately 28 million people in the U.S. have some degree of hearing impairment, with 25–40% of the population over 65 classified as hearing impaired. Elderly listeners often have an especially hard time understanding speech in noise or under distracting conditions [1]. The primary benefit of hearing aids is to restore hearing loss due to a loss of sensitivity by amplifying signal energy in one or multiple frequency bands, with optional dynamic compression and expansion schemes [2]. No attempts are made to perform prosodic or precise spectral modifications.

"Clear" (CLR) speech is a speaking style that speakers adopt to be understood correctly in a difficult communication environment. A number of studies have shown that CLR speech, as opposed to "conversational" (CNV) speech, has inherent intelligibility benefits. For example, CLR speech leads to significant improvements in intelligibility under noisy conditions for elderly listeners with sensorineural hearing loss [3, 4, 5].

Acoustic differences between CNV and CLR speech include many changes in prosodic and spectral cues [6, 7]. Affected prosodic cues include: an increased average fundamental frequency (F0) with increased variance; an increased consonant-vowel energy ratio, especially for stops; longer and more frequent pauses; lengthened vowels especially tense vowels (/i:/, /u/, />/, and /A/); and a significantly decreased speaking rate of 90–100 words per minute, as compared to 160–200 words per minute in CNV speech. Affected spectral cues include: lax vowels (/I/, /E/, /@/, /U/, and /&/) span a larger vowel space; formant frequencies approach their target; stop releases are more frequent. However, despite identifying acoustic differences little is known about the relationship between speech intelligibility and various acoustic features that are typical in CLR speech, i.e. it is unknown which features, and to what degree, make CLR speech more intelligible.

Signal processing algorithms have been developed that modify speech to approximate aspects of CLR speech, including decreasing the speaking rate, non-uniformly lengthening certain vowel or consonant durations, and enhancing the consonant-to-vowel energy ratios [8, 9, 10], with limited success. Our own preliminary experiments showed that manual tuning of prosody by experts does not improve intelligibility: Using 12 sentences and 18 listeners, two researchers modified natural sentences to create additional phrase breaks by manually altering F0, duration, and energy. The original sentences had an average intelligibility level of 78.1%, whereas the modified sentences had levels of 71.2% and 80.2% (not a statistically significant improvement). Uniform duration stretching resulted in a level of 76.0%.

Our long-term goal is to develop a model of the contribution of acoustic features to speech intelligibility. Applications of this model include (1) objective measures of speech intelligibility, (2) novel signal processing algorithms for hearing aids which can transform CNV speech into an approximation of CLR speech, and (3) algorithms for post-processing speech output from communications devices. Our short-term goal is to identify which features are responsible for improved intelligibility. We propose to achieve this goal by using a hybridization algorithm: Using parallel recordings of CNV and CLR speech, we replace a single feature, or a combination of features, of CNV speech with those extracted from CLR speech, thus creating hybrid speech. To identify which features contribute to improved intelligibility, we conduct perceptual experiments to evaluate the intelligibility of the original and hybrid sentences. We hypothesize that certain hybrid speech stimuli will have improved intelligibility, as compared to baseline CNV speech.

2. Speech Corpus

2.1. Text Material and Recording

We used the IEEE Harvard Psychoacoustic Sentences [11], which are syntactically and semantically normal (e. g. *His shirt was clean but one button was gone*). The sentences are crafted so that the various phonemes of English are present in accordance with their frequency of occurrence in the language.

One male, native speaker of American English recorded 70 sentences in the two styles CNV and CLR. When recording the CNV speech he was instructed to recite the material in a manner similar to everyday communication. When recording the CLR speech, he was instructed to speak as clearly as possible,

Name	Energy	F0	Duration	Spectrum	Phoneme	Non-speech
CNV	CNV	CNV	CNV	CNV	CNV	CNV
CLR-D	CNV	CNV	CLR	CNV	CNV	CNV
CLR-SP	CNV	CNV	CNV	CLR	CLR	CNV
CLR-DSP	CNV	CNV	CLR	CLR	CLR	CNV
CLR-EFN	CLR	CLR	CNV	CNV	CNV	CLR
CLR-EFDN	CLR	CLR	CLR	CNV	CNV	CLR
CLR	CLR	CLR	CLR	CLR	CLR	CLR

Table 1: Configurations governing the hybridization algorithm. Each configuration determines the source of six aspects of speech.

as if communicating with hearing-impaired or cognitively impaired listeners. For each sentence, we added manually verified phoneme labels and pitch marks. Phoneme statistics confirmed typical differences between the two speaking styles: CLR speech had more pauses and stop releases, while CNV speech had a larger number of reduced vowels.

2.2. Equalizing Loudness

Since increased loudness can contribute to speech intelligibility, it is important to eliminate loudness as a factor by equalizing the loudness of all sentences in the corpus. For each sentence, we considered all non-pausal speech segments and calculated the waveform peak and the loudness as defined by the A-weighted [12], root-mean-squared signal (rmsA). We then multiplied each waveform by varying gains so that all sentences had the same measured rmsA, while ensuring good resolution with adequate headroom (for later energy modifications).

2.3. Verification

In a first test, we intended to verify that all CNV sentences are in fact intelligible in the absence of background noise. Five young, normal-hearing listeners listened to all sentences (70 sentences total). A test administrator measured key word identification, and a sentence was scored as correct when five out of five words were identified correctly. The resulting sentence-level intelligibility rates were 99%. This confirmed that the conversationally spoken sentences were intelligible to the majority of listeners. We assumed that identification rates for the CLR sentences would be even higher.

In a second test, we compared speaking styles under the noise condition. We created stimuli by adding 12-talker babble noise [13] to the speech signals at variable signal-to-noise ratios (SNR). Sound pressure levels of the signal (without noise) were kept constant at 65 dBA (averaged over 10 seconds). Four young, normal-hearing listeners set the SNR using an "up-down" method with a 50% threshold [14]. We choose the 50% threshold in order to be in the middle of the perceptual performance curve, making the test as sensitive as possible. The procedure is as follows: Initially, the SNR is set to -3 dB. The first sentence is repeated at increasing SNR levels until the listener can obtain the correct response. After the first correct response, a different sentence is presented each time. The noise level is increased (SNR decreases) when the response is positive, and the noise level is decreased (SNR increases) when negative. Whenever the direction of the response is reversed (e. g. a positive response followed by a negative response), it is counted as one reversal. The increment or decrement of the SNR starts with a 2 dB step size, but after 3 reversals the step size is decreased to 1 dB. The test is continued until 8 reversals are accomplished. The final Speech Reception Threshold (SRT) is estimated by averaging SNR levels from reversals 3–8. As in the previous test, an administrator measured key word identification, and a sentence was scored as correct when five out of five words were identified correctly. The resulting SRT values were 3.5 dB and -0.7 dB for CNV and CLR speech. This shows that young listeners could more easily identify words spoken in the CLR style. Differences in speaking styles were statistically significant ($p = 0.05$), consistent with previous findings [3].

3. Hybridization Method

The purpose of hybridization is to obtain a speech waveform that combines aspects of CNV and CLR speech waveforms. Table 1 shows configurations that were used in the experiments (Section 4), by selecting a source for the following six aspects: energy trajectory, F0 trajectory, phoneme duration, short-term spectrum, phonemic content, and presence of non-speech sounds. The hybridization algorithm is implemented by first aligning the CLR phonetic sequence with the CNV one, then determining the desired hybrid phoneme sequence, and finally "parallelizing" the two original waveforms. Subsequently, a feature analysis is carried out on the parallelized (in terms of phonetic content, but not time-aligned) waveforms. Then, configuration-specific features of the CNV speech are replaced with those of the CLR speech to form features of hybrid speech. These hybrid features are then synthesized in the final synthesis step.

3.1. Phoneme Alignment and Waveform Parallelization

Hybridization necessitates that both CNV and CLR speech have the same phonetic content. Since this is not the case most of the time, we parallelize the phonetic content of both sequences. The sequences are often different because, even though the identical sentence was used, a speaker may pronounce the material differently, depending on the speaking style. As a first step, we align the phonetic sequences of both the CNV and the CLR speech. To accomplish the alignment, we use a phoneme feature table specifying numeric voicing, manner, place, and height features, with one 4-dimensional vector for each phoneme. Each phonetic symbol in the label sequences is assigned its associated feature vector, resulting in two feature matrices. Then, dynamic time warping is used to find an alignment path between the two matrices that yields the smallest distances between the corresponding phonetic features. The path is then expressed as a list of operations (substitution, insertion/deletion, no operation) and stored. This operations list was sometimes changed manually based on expert phonetic knowledge.

The desired hybrid phoneme sequence is dependent on the values of the "phoneme" and "non-speech" (such as pause and breath-noise) settings (see Table 1). Parallelization of the original waveforms to the hybrid phoneme sequence requires phoneme insertion and deletion operations. For phoneme inser-

CNV	CLR	Operation	Hybrid
b	b	-	b
j	j	-	j
u	u	-	u
-	tc	ins / del	tc
d_(	th	sub	d_(+ th
i:	i:	-	i:
-	.pau	ins / del	-

Table 2: Phoneme alignment operations and corresponding parallelization. The first two columns contain the CNV and the CLR speech phoneme sequence, after alignment, with the third column indicating the corresponding operation. The last column contains the hybrid phoneme sequence obtained when setting *phoneme*=CLR and *non-speech*=CNV, necessitating an insertion of a CLR /tc/ into the CNV speech, and a deletion of /.pau/ from the CLR speech.

tions, we extract the relevant portion from the corresponding alternative condition. As a result, no actual hybridization can take place in these regions. Phoneme substitutions are recorded in the parallelized phoneme string (e. g. /d_(+ th/), but no waveform operations are carried out. Table 2 shows the result of an example alignment and parallelization.

3.2. Feature Extraction

We extract acoustic features from CLR and CNV speech by performing a frame-by-frame, pitch-synchronous analysis. Frames of speech are defined by three neighboring (pitch or auxiliary) marks, representing their leftmost, center, and rightmost positions, thus spanning two periods of speech during voiced regions. Marks are obtained by segmenting the speech signal using pitch marks and additional auxiliary marks, which are created in unvoiced regions at regular intervals. An individual frame is considered to represent the short-term spectrum at the time of the frame center. To obtain an energy value, we apply a non-symmetric Hanning window to a single frame, and then calculate the rmsA of the signal. F0 is obtained by inverting the difference between pitch marks. Finally, speech durations are specified at the phoneme level and can be directly derived from the labels in the speech corpus.

3.3. Feature Replacement and Synthesis

In this final step, we replace a specified single feature, or a group of features, of the CNV speech with the same type of features extracted from the CLR speech. The resulting hybrid features are then synthesized by pitch-synchronous, overlap-add, residual-excited, linear predictive coefficient (LPC) synthesis that implement energy, F0, and duration changes using the specified short-term speech signal frames, thus creating the hybrid speech.

4. Perceptual Tests

We recruited listeners whose first language was American English. They were instructed to repeat the sentences they heard as best as they could. Stimuli were played over circumaural headphones (Sennheiser HD 280 Pro), binaurally. Similar to Section 2.3, speech signals were delivered at 65 dBA, with added 12-talker babble noise. For each listener, we first measured their SRT, and then used the resulting SNR values to measure intelligibility levels for 48 sentences. Sentences were considered correct

Experiment	Condition	Intelligibility (SD)	DOC
1	CNV	66 (12)	0
	CLR-EFDN	60 (18)	-43
	CLR-SP	55 (21)	-79
	CLR	80 (17) $*$	100
2	CNV	69 (12)	0
	CLR-EFN	61 (20)	-32
	CLR-DSP	92 (9) $*$	92
	CLR	94 (12) $*$	100
3	CNV	64 (20)	0
	CLR-DSP$_2$	82 (15) $*$	72
	CLR-D$_2$	74 (18)	40
	CLR-SP$_2$	75 (14) $*$	44
	CLR-SP	57 (27)	-28
	CLR	89 (10) $*$	100

Table 3: Intelligibility rates in percent and standard deviations in parentheses. The degree of contribution (DOC) in percent is shown in the right column. Results marked with an asterisk are significantly different ($p = 0.05$) as compared to the CNV condition of that experiment.

if listeners identified four out of five key words.

4.1. Experiment 1

In the first experiment, we were interested in measuring varying intelligibility levels among the CNV, CLR-EFDN (the "prosodic" group), CLR-SP (the "spectral" group), and CLR configurations. Eight listeners between 65 and 75 (average age: 69) participated in the study. All subjects had hearing sensitivity less than 35 dBHL in the range of 250–4000 Hz.

Average SRT value was 1.47 dB. Measured intelligibility levels and the degree of contribution (DOC) are shown in Table 3. We define DOC as a value given by the ratio $D = (I_{HYB} - I_{CNV})/(I_{CLR} - I_{CNV})$, where I represents intelligibility levels and the subscript refers to a specific condition. The intelligibility of CLR speech was significantly improved over CNV speech. However, hybridization with any materials did not yield improvements; in fact, intelligibility levels were slightly worse, although not always significantly so. We posited three hypotheses that can explain these results: (H1) Phoneme durations can not be separated from spectral features, (H2) speech processing artifacts of the hybridization algorithm degrade the speech signal, and (H3) elderly listeners perform differently from young listeners. The following experiments are designed to test these hypotheses.

4.2. Experiment 2

In this experiment we used configurations CNV, CLR-EFN, CLR-DSP, and CLR. In contrast with Experiment 1, the source of phoneme duration and spectral features are matched in order to test H1. Eight young (average age: 28), normal-hearing listeners participated, as defined by a hearing sensitivity of less than 25 dBHL in the range of 250–4000 Hz.

Average SRT value was 0.58 dB. The SRT value is lower than for elderly listeners, as expected. Levels of the CLR-DSP hybrid proved to be significantly above that of the CNV speech and the degree of contribution was 92%, supporting H1. In order to confirm H1, we examined CLR-D, CLR-SP, CLR-DSP with young normal-hearing listeners in the next experiment. H2 was not addressed in Experiment 2. The intelligibility difference be-

tween CNV and CLR speech for young normal-hearing listeners was much larger than for elderly listeners, supporting hypothesis H3. Finally, CLR-EFN gave no improvements (see Table 3).

4.3. Experiment 3

To test hypothesis H2, we addressed possible sources of artifacts in the hybridization algorithm. The new implementation included the following changes:

1. Allowing very low F0 due to glottalization in voiced sounds.
2. Preventing frame duplication due to pitch or duration modification during impulsive speech regions such as bursts.
3. Preventing frame duplication near voiced-unvoiced transitions.
4. Smoothly fading in and out of phoneme insertions or deletions, as required by the parallelization step (Section 3.1).

The final experiment used configurations CNV, CLR-DSP$_2$, CLR-D$_2$, CLR-SP$_2$, CLR-SP, and CLR, where the subscript "2" indicates that the condition was produced by the new implementation. Eight young (average age: 29), normal-hearing listeners participated.

Average SRT value was -0.24 dB. Conditions CLR-DSP$_2$ and CLR-SP$_2$ were both significantly different from CNV (and CLR-D$_2$ was nearly so with $p = 0.07$). Spectral features separating from the source of duration yielded improvement; therefore, hypothesis H1 is not confirmed. Also, the difference between CLR-SP$_2$ and CLR-SP was significant, indicating that the new implementation yielded a significant improvement, confirming H2. Similar to Experiment 2, the intelligibility of CLR speech was significantly better than that of CNV speech for young listeners, again supporting hypothesis H3 (see Table 3).

We also performed an additional quality test using young, normal-hearing listeners. For each of the four implementation improvements, we presented pairs of stimuli with a particular improvement turned on and off. Listeners were asked to rate the quality difference using a scale of -2 (much worse), -1 (worse), 0 (about the same), +1 (better), and +2 (much better). Averaging the scores results in a comparative mean opinion score (CMOS), which indicated significant improvements for implementation changes 2 (+0.6), and 4 (+0.4), but a worsening for change 1 (-0.5).

5. Conclusion

First, we confirmed that CLR speech has consistently higher intelligibility than CNV speech in a series of perceptual experiments. Intelligibility differences were larger for young listeners than for elderly listeners. This suggests that we need to obtain larger differences between CNV and CLR speech using elderly listeners in order to improve intelligibility of CNV speech in the future.

Stimuli were created using a hybridization algorithm that combined acoustic features of CNV and CLR speech. The results showed that hybrid speech, which has acoustic features from CNV and CLR speech, can improve speech intelligibility over the baseline CNV speech. We conclude that a combination of phoneme duration, phoneme identity and spectral features, as well as a combination of phoneme identity and spectral features contribute to high intelligibility in CLR speech.

It is imperative to address any speech processing artifacts in the hybridization algorithm, and that improvements of intelligibility from CNV speech would depend on the combinations of acoustic features used, the quality of the algorithm implementation, and the age of the listeners. The hybridization algorithm is equivalent to modifying CNV speech with an "oracle" mapping function, thus simulating maximum performance levels of an automatic modification system.

6. References

[1] T. A. Salthouse, "The processing-speech theory of adult age differences in cognition," *Psychological Review*, vol. 103, no. 3, pp. 403–428, 1996.

[2] H. Dillon, *Hearing aids*, Thieme, New York, 2001.

[3] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech.," *Journal of Speech and Hearing Research*, vol. 28, pp. 96–103, 1985.

[4] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," *Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1581–1592, 1994.

[5] S. H. Ferguson and D. Kewley-Port, "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 112, no. 1, pp. 259–271, 2002.

[6] 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.

[7] J. C. Krause and L. D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *Journal of the Acoustical Society of America*, vol. 15, no. 1, pp. 362–378, 2004.

[8] S. Gordon-Salant, "Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing," *Journal of the Acoustical Society of America*, vol. 82, no. 6, pp. 1599–1607, 1986.

[9] Y. Nejime and B. C. J. Moore, "Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss," *Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 572–576, 1998.

[10] R. M. Uchanski, S. S. Choi, L. D. Braida, C. M. Reed, and N. I. Durlach, "Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate.," *Journal of Speech and Hearing Research*, vol. 39, pp. 494–509, 1996.

[11] 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 Transactions on Audio Electroacoustics*, vol. 17, pp. 227–246, 1969.

[12] International Electrotechnical Commission, *Electroacoustics-sound level meters-part 1: Specifications, 61672*, 2002.

[13] R. C. Bilger and J. M. Nuetzel, "Standardization of a test of speech perception in noise," *Journal of Speech and Hearing Research*, vol. 27, pp. 32–48, 1984.

[14] H. Levitt, "Transformed up-down methods in psychoacoustics," *The Journal of the Acoustical Society of America*, vol. 49, no. 2, pp. 467–477, 1971.