Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility

Alexander Kain, Akiko Amano-Kusumoto, and John-Paul Hosom
Center for Spoken Language Understanding (CSLU) at the OGI School of Science and Engineering, Oregon Health and Science University (OHSU), 20000 NW Walker Road, Beaverton, Oregon 97006

(Received 1 June 2007; revised 1 July 2008; accepted 7 July 2008)

Speakers naturally adopt a special "clear" (CLR) speaking style in order to be better understood by listeners who are moderately impaired in their ability to understand speech due to a hearing impairment, the presence of background noise, or both. In contrast, speech intended for nonimpaired listeners in quiet environments is referred to as "conversational" (CNV). Studies have shown that the intelligibility of CLR speech is usually higher than that of CNV speech in adverse circumstances. It is not known which individual acoustic features or combinations of features cause the higher intelligibility of CLR speech. The objective of this study is to determine the contribution of some acoustic features to intelligibility for a single speaker. The proposed method creates "hybrid" (HYB) speech stimuli that selectively combine acoustic features of one sentence spoken in the CNV and CLR styles. The intelligibility of these stimuli is then measured in perceptual tests, using 96 phonetically balanced sentences. Results for one speaker show significant sentence-level intelligibility improvements over CNV speech when replacing certain combinations of short-term spectra, phoneme identities, and phoneme durations of CNV speech with those from CLR speech, but no improvements for combinations involving fundamental frequency, energy, or nonspeech events (pauses). © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2967844]

PACS number(s): 43.71.Gv, 43.72.Ja, 43.71.Es [DOS] Pages: 2308–2319

I. INTRODUCTION

Approximately 28×10^6 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 (Cruickshanks *et al.*, 1998). Elderly listeners often have an especially difficult time understanding speech in noise or under distracting conditions (Salthouse, 1996). 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, with optional dynamic compression and expansion (Dillon, 2001). 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" (e.g., Picheny *et al.*, 1985). In contrast, speech intended for a normal-hearing listener in a quiet environment is commonly referred to as "conversational." These two styles of speaking will be referred to as CLR and CNV speech, respectively. 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 (Picheny *et al.*, 1985; Payton *et al.*, 1994; Schum, 1996; Helfer, 1998; Bradlow and Bent, 2002; Ferguson and Kewley-Port, 2002; Krause and Braida, 2002; Bradlow *et al.*, 2003; Ferguson, 2004; Liu *et al.*, 2004).

Previous research has examined acoustic differences between CNV and CLR speech. The following *prosodic* features have been noted to distinguish CLR speech from CNV speech: (1) The fundamental frequency (F0) was typically increased in range and mean (Picheny *et al.*, 1986; Bradlow *et al.*, 2003; Krause and Braida, 2004); (2) the consonant-vowel energy ratio (CVR) was increased (energies of the consonants were greater in CLR speech), particularly for stops and affricates (Bradlow *et al.*, 2003); however, other researchers found an increased CVR for affricates only (Krause and Braida, 2004); (3) phoneme durations were prolonged, especially in the tense vowels /i/, /u/, /ɑ/, and /ɔ/ (Picheny *et al.*, 1986; Ferguson and Kewley-Port, 2002); (4) pauses were longer in duration and occurred more frequently (Picheny *et al.*, 1986; Bradlow *et al.*, 2003); and (5) the speaking rate was significantly decreased from 160–200 words/min to 90–100 words/min, which was at least partially the result of longer vowels in CLR speech (Picheny *et al.*, 1986; Krause and Braida, 2004).

The following *spectral* features have been noted to distinguish CLR speech from CNV speech: (1) Vowel formant frequencies showed expanded vowel spaces for lax vowels, (Picheny *et al.*, 1986), (2) long-term spectra had increased energies at higher frequencies (1000–3150 Hz) (Krause and Braida, 2004), (3) alveolar flaps occurred less often and consonant stops (e.g., /t/) tended to be released with following aspiration (Picheny *et al.*, 1986; Krause and Braida, 2004; Bradlow *et al.*, 2003); and (4) four out of five speakers

 0001-4966/2008/124(4)/2308/12/$23.00

exhibited increased modulation indices for low modulation frequencies up to 3–4 Hz (Krause and Braida, 2004).

Different speakers appear to employ different strategies to produce CLR speech, resulting in significantly different vowel intelligibility (Ferguson and Kewley-Port, 2002; Ferguson, 2004). In the same study, the authors found that not all speakers can improve their intelligibility beyond their CNV speech level. The least intelligible speakers produced CLR speech with (1) shorter durations, (2) the least differentiated vowel spaces as measured by the first two formants, (3) minimal cues for consonantal contrasts, and (4) the most varied amplitude of stressed vowels (Bond and Moore, 1994). Speakers with larger vowel spaces were generally more intelligible than speakers with reduced vowel spaces (Bradlow *et al.*, 1996). The relative intelligibility of an individual speaker has been shown to be consistent across listener age groups (7–8 year olds, 11–12 year olds, and adults with mean age of 29.9 years) (Hazan and Markham, 2004).

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 (Pichora-Fuller *et al.*, 1995). For example, elderly listeners have more difficulty than younger listeners understanding rapid speech due to decreased auditory processing capabilities (e.g., Gordon-Salant and Fitzgibons, 2001) or reduced working memory capacity (Wingfield *et al.*, 2005). Motivated by these findings, researchers have developed signal-processing algorithms to increase the intelligibility of speech independent of amplification. Modifications to the speech waveform include decreasing the rate of speech by inserting pauses (Liu and Zeng, 2006), modifying phoneme durations (Nejime and Moore, 1998; Gordon-Salant, 1986; Uchanski *et al.*, 1996), and enhancing the consonant-to-vowel energy ratios (Gordon-Salant, 1986, 1987; Hazan and Simpson, 1998). Of these, only one study showed a statistically significant intelligibility increase of 4.2% at the sentence level, by amplifying the energy of specific consonants (Hazan and Simpson, 1998). Other studies did not report statistically significant improvements at the sentence level; this relative lack of success was sometimes conjectured to be caused by signal-processing artifacts (e.g., Uchanski *et al.*, 1996), although the manipulation of isolated features may also have been a contributing factor, as interactions among features are common. Ultimately, the causal relationship between sets of acoustic features and speech intelligibility, using sentence-level speech materials, is not known.

The long-term goal of this research effort is to develop a model that quantifies the cause-and-effect relationship between acoustic features of speech and speech intelligibility. Such a model may have applications in novel signal processing algorithms for hearing aids and other assistive listening devices that transform CNV speech into a closer approximation of CLR speech, for postprocessing speech output from general-purpose communication devices (e.g., phones and video playback devices), and for objective measures of speech intelligibility.

This article reports on initial experiments that measured the degree of contribution (DOC) of six high-level acoustic features to intelligibility. It is not the intent of this work to provide a complete list of all relevant feature combinations with their respective degrees of contribution, but to demonstrate that it is possible to determine the significance of specific feature combinations to intelligibility. Similar to previous approaches, in which phoneme durations from CLR speech were applied to CNV speech (Uchanski *et al.*, 1996), or in which the temporal envelope from CLR speech was applied to CNV speech using "chimerized" speech (Liu and Zeng, 2006), the present work applies certain CLR features to CNV speech. 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 allows us to identify not only which acoustic features contribute to intelligibility but also their DOC. The DOC is subject to listener variation; because of the relatively small number of listeners used in the following experiments, DOC values are not precise metrics but are provided to give a sense of how much impact a feature has. It was hypothesized that certain subsets of acoustic features of CLR speech contribute significantly more to speech intelligibility than others.

II. SPEECH CORPUS

A. Text material and recording

Two types of text material were used in this study, referred to as material A and material B.

Material A. This material consisted of 70 sentences from the set of IEEE Harvard Psychoacoustic Sentences (Rothauser *et al.*, 1969), which are phonetically balanced, syntactically, and semantically normal sentences, with each sentence containing five keywords (e.g., His *shirt* was *clean* but *one button* was *gone*).

Material B. This material consisted of 70 syntactically correct, but semantically anomalous sentences (e.g., They *slide far across* the *tiny clock*), created by randomizing and exchanging words and grammar structures from material A (Vaughan *et al.*, 2006). Using identical words and sentence lengths in both materials allowed for direct comparisons between experimental results.

It was expected that material A would be easier to understand because the availability of semantic context is an important factor contributing to speech intelligibility (Gordon-Salant and Fitzgibbons, 1997), but these sentences are still not as predictable as everyday speech (Rothauser *et al.*, 1969).

One male, a native speaker of American English with no professional training in public speaking, was recruited as a speaker (J.P.H.). First, he recorded the 140 sentences of materials A and B spoken in the CNV speaking style, followed by the same sentences spoken in the CLR speaking style. When recording CNV speech, he was instructed to speak in

TABLE I. Example configurations governing the hybridization algorithm. Each configuration determines a speech style (CNV or CLR) as source for six acoustic features of speech.

Name	Energy	F0	Duration	Spectrum	Phoneme	Non speech
CNV	CNV	CNV	CNV	CNV	CNV	CNV
HYB-D	CNV	CNV	CLR	CNV	CNV	CNV
HYB-EFN	CLR	CLR	CNV	CNV	CNV	CLR
CLR	CLR	CLR	CLR	CLR	CLR	CLR

the way that he uses to communicate in his daily life. When recording CLR speech, he was instructed to speak clearly as he would when communicating with hearing-impaired listeners.

The recording was carried out in a sound-treated booth (Whisperroom MDL4260) located inside a control room. Recordings were made using a head-mounted close-talking microphone (AKG HSC200), positioned approximately 5 cm and off axis from the speaker's mouth. The speaker recorded the materials at his own pace by operating a computer program. A technician listened to each sentence and the speaker was asked to record a sentence again when pronunciation or style were not satisfactory. The speech signals were captured and stored digitally at a sampling rate of 22.05 kHz with 16 bit resolution.

Sentences were partitioned into two groups, based on an informal characterization of the degree of difference between renditions of the two styles. For each material, the 48 sentences with apparently larger intelligibility differences (partition LD) were used for intelligibility testing (see Sec. IV D), while the remaining 22 sentences with apparently smaller differences (partition SD) were used for setting noise levels for those tests (see Sec. IV B).

B. Additional annotation

Time-aligned phoneme label and "pitch mark" (also known as glottal closure instant) annotation was added to the corpus of speech waveform recordings. Labels included nonspeech events, almost always occurring in the form of pauses, but also in the form of breath noise, lip smack, and tongue click. This additional annotation was required for processing the speech waveforms as part of the hybridization algorithm (see Sec. III). Initial estimates of phoneme identities and boundaries in each waveform were obtained using an existing forced-alignment system (Hosom, 2002). To create an initial estimate of the pitch marks in each waveform, a standard software package was used (Boersma, 1993). Then, a trained labeler checked and adjusted pitch marks as well as phoneme identities and boundaries manually. The entire corpus, including waveforms, phoneme labels, and pitch marks, is available through the CSLU corpus distribution mechanism (Center for Spoken Language Understanding, 2007).

C. Corpus characteristics

The phonetic and acoustic characteristics of the CNV and CLR sentences were analyzed using partition LD.

1. Phonetic characteristics

Phoneme statistics confirmed typical differences between the two speaking styles; for example, CLR speech had more pauses and stop releases, while CNV speech had a larger number of reduced vowels. A phonetic alignment (see Sec. III B 1) of partition LD of materials A and B (9090 labels in 96 sentences) resulted in 109 labels from CLR speech being considered as insertions into CNV speech (among them 38 pauses and 29 unvoiced plosive releases), while 20 labels from CNV speech were considered as insertions into CLR speech (8 of which were pauses). This corresponds to approximately 0.9 and 0.5 insertion/deletion operations per sentence, for phonemes and nonspeech events, respectively. Phoneme substitutions (e.g., /u/ versus /ʉ/) occurred 150 times.

2. Acoustic characteristics

An acoustic analysis showed that, for CNV speech, the mean and standard deviation (SD) of F0 were 105.0 and 15.4 Hz, respectively, whereas CLR speech had mean and SD values of 106.2 and 17.6 Hz, respectively. Using a frame-by-frame root-mean-squared energy measure, it was found that the energy of CLR speech (excluding nonspeech sounds) was 0.9 dB above that of CNV speech (1.1 dB for vowels only), and the SD of CLR speech energy was approximately 30% larger than the SD of CNV speech. The average CVRs were measured and determined to be −7.2 dB in CNV speech and −7.0 dB in CLR speech for plosives, and −6.9 dB in CNV speech and −7.1 dB in CLR speech for affricates. Spectrally, the vowels of CLR speech had an approximately constant energy increase of 2 dB in frequencies above 1800 Hz. Finally, the average duration of interword pauses was 33 ms in CNV speech and 62 ms in CLR speech.

III. HYBRIDIZATION ALGORITHM

The purpose of the hybridization algorithm was to obtain a speech waveform that combines acoustic features of the CLR and CNV speech waveforms. Table I shows example feature configurations and corresponding stimulus conditions (see Sec. IV D). Configurations involve the following six high-level acoustic features: energy trajectory (E), F0 trajectory (F), phoneme durations (D), short-term spectra (S), phoneme sequence (P), and presence of nonspeech sounds such as pause, breath noise, lip smack, and tongue click (N). Configurations are named according to the acoustic features that are taken from CLR speech; for example, HYB-EFN represents a HYB speech waveform whose fea-

tures consist of energy, F0, and nonspeech events from CLR speech (the ordering of features is immaterial), while all other features are taken from CNV speech. Experiments 1–3 (see Sec. IV) evaluated the intelligibility of these HYB speech stimuli to determine the DOC of particular sets of acoustic features used from CLR speech, as compared to the baseline CNV speech.

A. Normalizing levels

Because increased loudness can contribute to speech intelligibility, it was important to minimize sentence-level loudness differences by normalizing all sentences (both CLR and CNV styles, and both text materials) in the corpus, using the following procedure: Speech signal waveforms were first filtered with an A-weighted filter (International Electrotechnical Commission, 2002), and levels were calculated by averaging frame-by-frame (using frame durations on the order of 10 ms) root-mean-squared (rmsA) sample values of nonpausal (as determined by the phoneme alignment described in Sec. II B) portions of the speech. Each waveform was then multiplied by an appropriate gain factor so that the resulting waveforms all had the same average rmsA value, while at the same time ensuring good resolution with sufficient headroom (e.g., setting the global sample peak to 80% of the maximum absolute sample value). This allowed for possible energy increases during hybridization.

B. Algorithm implementation

The hybridization algorithm was implemented by first aligning the CLR phonetic sequence with the CNV phonetic sequence (step 1) and then "parallelizing" the two original waveforms in terms of phonetic content (step 2). Next, a speech analysis was carried out on the parallelized waveforms, consisting of first computing auxiliary marks (step 3) and then extracting acoustic features (step 4). Next, configuration-specific features of CLR speech were combined with complementary features of CNV speech to form HYB features (step 5) and finally synthesized to create the HYB speech waveform (step 6). Each step is now described in more detail.

1. Alignment of phoneme sequences

The phoneme sequences of CNV and CLR speech are occasionally different because, even when identical sentences are used, a speaker may pronounce the material differently, depending on the speaking style (see Sec. II C 1). However, the feature combination (step 5) required that the input speech signals have a one-to-one mapping between phoneme sequences; therefore, phoneme-sequence pairs of each sentence produced in the CNV and CLR styles needed to be aligned.

To accomplish the alignment, a phoneme feature table was created, specifying voicing (1–5: from fully voiced to voiceless, e.g., 1: vowels, 3: voiced fricatives), manner (0–10: from highest to lowest sonority, e.g., 3: glides, 6: fricatives), place (1–8: from front to back, e.g., 1: bilabial, 8: velar), and height (1–10: from lowest to highest tongue height, e.g., 1:/a/, 4:/u/) features), with one four-dimensional vector for each phoneme (e.g., /i/: [1,2,4,4,], /k/: [4,7,8,7]). Each phonetic symbol in both label sequences was assigned its associated feature vector, resulting in two feature matrices. Then, dynamic time warping (Rabiner and Juang, 1993) was used to find an optimal alignment path between the two matrices that resulted in the minimum Euclidean distance between the corresponding phonetic features, while observing sensible local constraints. 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 (or a deletion, depending on the point of view). The alignment path was then automatically converted to a list of operations (insertion/deletion for one-to-many mappings, no change for perfect or best matches) and stored. This operation list was sometimes changed manually based on phonetic knowledge.

TABLE II. An example of the phoneme alignment operations and corresponding parallelization for a HYB-P configuration. 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 *Nonspeech*=CNV, necessitating an insertion of the CLR plosive release /$d^{(\urcorner)}$/ into the CNV speech, and a deletion of the pause /(.)/ from the CLR speech.

CNV	CLR	Operation	HYB
b	b	⋯	b
ɪ	i	⋯	ɪ;i
s	s	⋯	s
aɪ	aɪ	⋯	aɪ
$d^{(\urcorner)}$	$d^{(\urcorner)}$	⋯	$d^{(\urcorner)}$
⋯	$d^{(\urcorner)}$	ins	$d^{(\urcorner)}$
⋯	(.)	del	⋯

2. Parallelization of waveforms

The final HYB phoneme sequence was dependent on the values of the phoneme (P) and nonspeech (N) configuration settings (see Table I). Therefore, the original waveforms were modified to implement phoneme deletions or insertions, which may occur in the CNV speech, the CLR speech, or both.

For phoneme insertions, the relevant portion was extracted from the corresponding alternative style and inserted without any time-domain cross-fade. (This technique was later modified, see Sec. IV E.) Phoneme deletions were carried out by removing the relevant portion from the waveform and concatenating the two remaining sections together. Because an inserted waveform segment resulted in identical waveforms in that region for both styles, hybridization was not possible in those regions.

Table II shows the result of an example phoneme alignment and the resulting parallelization, given a HYB-P configuration, which means that phonemes were taken from CLR speech, but that nonspeech events were taken from CNV speech. In this example, the CNV speech phoneme /ɪ/ was aligned with the CLR speech phoneme /i/, which was considered a substitution (noted in the HYB phoneme se-

quence as /ɪ; i/); hybridization can proceed normally here. The waveform associated with the /$d^{(\ulcorner)}$/ phoneme label from the CLR speech was copied and inserted into the CNV speech. (Note that in this work the symbol /$d^{(\urcorner)}$/ denotes /d/ closure and /$d^{(\ulcorner)}$/ denotes the released part of the plosive.) As a result, because both speaking styles contained identical waveforms for this phoneme, and thus all acoustic features were identical, hybridization could not take place in the /$d^{(\ulcorner)}$/ phoneme. Finally, the waveform associated with the (interword) pause /(.)/ was deleted from the CLR speech, in order to satisfy the requirement of following the nonspeech pattern of the CNV speech.

3. Computing auxiliary marks

The hybridization algorithm processed speech by performing computations on short successive segments of speech, known as frame-by-frame processing. The extent of each frame was defined by the location of consecutive "time marks." In voiced regions, these time marks were defined by pitch marks (the instants of glottal closure); in unvoiced regions, time marks were defined by "auxiliary marks," which needed to be created to keep the frame size appropriately short. Whenever the distance between two pitch marks was more than 16 ms (or, equivalently, when F0 was lower than 62.5 Hz), the region was considered unvoiced. Auxiliary marks were placed inside unvoiced regions at intervals of approximately 10 ms. Care was taken that auxiliary marks and pitch marks were not closer than 10 ms to each other, thus avoiding very short frames that could lead to problems during the feature extraction step.

4. Feature extraction

Acoustic features from CNV and CLR speech were extracted by performing a frame-by-frame pitch-synchronous analysis. Consecutive overlapping analysis frames were defined by three consecutive time marks spanning two periods of speech. The short-term spectrum (S) was represented by storing the analysis frame's waveform directly. To obtain an energy trajectory (E), a Hanning window was applied to each analysis frame, and then the corresponding rmsA values were calculated (see Sec. III A). F0 trajectory values (F) were obtained by inverting the differences between pitch marks. These features were considered to represent the speech waveform at the time of their respective analysis frame centers. Finally, speech durations (D) were specified at the phoneme level and could be directly derived from the labels in the speech corpus.

5. Feature combination

In preparation for the final synthesis step, HYB features were formed from a particular subset of CLR features and from the complementary subset of CNV features, as specified by the desired configuration (see Table I). For example, the configuration HYB-DSP specified that the CLR speech was taken as the source for phoneme durations, spectral information, and phoneme sequence, and that the complementary features, namely, energy trajectory, F0 trajectory, and nonspeech sequences, were taken from CNV speech.

6. Speech synthesis

The HYB features created in the previous step were used to synthesize a HYB speech waveform using a pitch-synchronous, overlap-add, residual-excited, linear predictive coefficient (LPC) synthesis method (similar to Taylor *et al.*, 1998). The method controlled energy trajectories, F0 trajectories, and phoneme durations using short frames of speech, which were constructed from three consecutive time marks (spanning the waveform of two speech periods).

F0 modification of a frame of speech was implemented by first pre-emphasizing the speech and then calculating 24th-order LPC coefficients (using the autocorrelation method) and corresponding LPC residuals. The length of the residual was changed to match the new speech period, by adding zeros to both ends symmetrically when decreasing F0 or shortening the residual for increasing F0. Finally, the original LPC filter was excited with the modified residual. Duration modification was implemented by repetition (for time expansion) or deletion (for time compression) of individual frames. Frames were selected by linear sampling throughout the duration of the phoneme. Energy modification was carried out by first calculating gain factors for each frame, given by the specified energy divided by the existing energy of that frame. The gain factor trajectory was first filtered by a tenth-order median filter and then smoothed using a zero-phase low-pass filter (five-point symmetric, normalized Hanning window). The intent of these filters was to preserve existing frame-to-frame energy fluctuations, while implementing the desired longer-term energy changes. Finally, the waveforms of each frame were scaled by their associated gain factor.

For the overlap-add operation, each frame was windowed and the second half of one frame was added to the first half of the next frame. The window was an asymmetric trapezoidal given by

$$w(t|l,r)=\begin{cases}0 & \text{if } t=1\cdots\dfrac{l}{2}\\ \dfrac{t-l/2}{l/2+1} & \text{if } t=\left(\dfrac{l}{2}+1\right)\cdots l\\ 1, & \text{if } t=(l+1)\cdots\left(l+\dfrac{r}{2}\right)\\ \dfrac{t-l-r/2}{r/2+1} & \text{if } t=\left(l+\dfrac{r}{2}+1\right)\cdots(l+r),\end{cases} \tag{1}$$

where $t=1\cdots l+r$ represents time and l and r are the lengths of the left and right speech periods of a frame. The advantage of using this trapezoidal window is that it avoids LPC filter startup artifacts by discarding initial sample values, and that it continues to "ring" beyond the original excitation, desirable features when using shortened or lengthened (zero-padded) residuals.

The hybridization algorithm can be seen as a LPC resynthesis of either CNV or CLR speech, while optionally modifying energy, F0, and phoneme durations, as well as inserting or deleting phonemes. For example, the condition HYB-SPN is equivalent to a resynthesis of CLR speech with energy, F0,

and phoneme durations modified to match the corresponding acoustic features extracted from the CNV speech. Conversely, this configuration can be thought of as a resynthesis of CNV speech where frame waveforms have been replaced with the corresponding and appropriately selected and modified frame waveforms from CLR speech, in addition to inserting or deleting waveform segments as required by the CLR speech labels during the parallelization process.

The hybridization algorithm shares some characteristics with voice transformation (VT) algorithms. During training, VT algorithms estimate parameters of a mapping function that predicts target features from time-aligned source features, extracted from a target and a source speaker, respectively. During transformation, a new source speaker utterance is analyzed, its features transformed by the trained mapping function, and a new speech signal is synthesized with characteristics of the target speaker. Applications include speaker modification (Kain, 2001), improving the intelligibility of dysarthric speech (Kain *et al.*, 2007), and speech morphing (Abe, 1996). However, the current work is distinct from VT algorithms in that its goal is the study of feature selection and the accompanying changes in intelligibility. Moreover, as the CLR speech features are known, there is no training function and no mapping in the current work.

IV. PERCEPTUAL EXPERIMENTS

Three experiments were conducted to test intelligibility differences as a function of speech material and hybrid configurations. The conditions of the second and third experiments depended on the results from the previous experiment, and so each experiment employed a unique set of listeners and should be considered in sequence. Intelligibility results between experiments are not directly comparable because the set of listeners varied between experiments. Because of this variability between listener groups, we restrict our comparison of results, other than noting significance, to within-experiment results.

Each perceptual test was carried out on a fanless computer with an external sound card (M-Audio USB Duo). To ensure that experimental stimuli were administered at controlled fixed sound pressure levels while minimizing ambient sounds, stimuli were presented through headphones in a quiet room. The speech signal was calibrated acoustically at 65 dBA, as measured in a custom-made coupler with a sound-level meter (Brüel and Kjær Type 2238) attached to a condenser microphone (Brüel and Kjær Type 4133). The signal level was averaged over 10 s of continuous speech, excluding pauses. Throughout the study, whenever noise was added to speech, the energy of the speech signal was kept constant, while the energy of the noise was varied. The noise consisted of near-field microphone recordings of 12 speakers that were combined digitally, called "babble" noise (Bilger and Nuetzel, 1984).

A. Listeners

Men and women aged 18–39 with self-reported normal hearing participated in the experiments described in this section. The first language of all listeners was North-American English. Every perceptual test was carried out using unique listeners.

B. Obtaining SNR50% values

The term SNR50% (Chang *et al.*, 2006) refers to that signal-to-noise ratio (SNR) at which listeners can correctly identify key words in the presence of noise 50% of the time. Experiments were carried out with the noise level set to each listener's individual SNR50% level, to help normalize for differences in hearing performance between listeners.

The approach to estimate SNR50% levels was based on the "up-down" method (Levitt, 1971). The details of the procedure were as follows: Initially, the SNR was set to −3 dB. The first sentence was repeated at increasing SNR levels until the listener could obtain the correct response. Four keywords were required to be recalled correctly for a sentence to be considered correct. After the first correct response, a different sentence was presented each time. The noise level was increased (SNR decreased) when the listener's response was correct, and the noise level was decreased (SNR increased) when the response was incorrect. A correct response followed by an incorrect response, or vice versa, was counted as one reversal. The procedure was continued until eight reversals took place. Finally, the SNR50% value was computed by averaging SNR levels from reversals 3 to 8.

C. Speech corpus verification

In a first preliminary test, the intention was to verify that all (unmodified) CNV sentences (in both SD and LD partitions) were, in fact, intelligible in the absence of background noise. Four listeners listened to all sentences from materials A and B (140 sentences total). A test administrator measured key word identification, scoring a sentence as correct when five out of five words were identified correctly. The resulting sentence-level intelligibility rates were 98.57% and 92.50% for materials A and B, respectively. This confirmed that the CNV sentences were generally intelligible. It was assumed that the intelligibility of CLR sentences would be even higher.

In a second preliminary test, speaking styles and material types were compared by determining SNR50% values, using the procedure described in Sec. IV B. Eight listeners first listened to CNV sentences of materials A and B from the SD partition, and then listened to CNV and CLR sentences of both material types from the LD partition in a Latin Square design. The size of the SD partition did not allow testing of two styles; therefore, only the CNV style was evaluated, as in the initial SNR50% tests of the main intelligibility experiments. SNR50% values for each of these six conditions were determined separately. An administrator measured key word identification, scoring a sentence as correct when four out of five words were identified correctly. Results are shown in the first two rows of Table III. The data of the LD partition were submitted to a 2×2 (speaking styles CLR and CNV, speech materials A and B) analysis of variance. There were significant ($\alpha=0.05$) main effects of speaking style ($F(1,28)=7.32$, $p=0.012$) and sentence mate-

TABLE III. Listeners' average SNR50% levels in dB, with SDs in parentheses, for materials A and B in partitions SD and LD. Results are from the speech corpus verification (V) test, as well as from the three intelligibility experiments (E).

Test	A-SD	B-SD	A-LD	B-LD
V. CNV	−0.71(1.18)	−0.39(1.68)	−2.52(0.84)	−1.53(1.50)
V. CLR	n/a	n/a	−4.26(1.25)	−2.54(1.94)
E. 1 CNV	0.58(0.88)	1.81(1.78)	n/a	n/a
E. 2 CNV	−0.24(1.11)	−0.07(1.39)	n/a	n/a
E. 3 CNV	0.22(1.44)	0.30(1.74)	n/a	n/a

rial ($F(1,28)=7.1$, $p=0.013$). Results between the partitions cannot be compared because the SD partition was always tested first. No significant interactions between speaking style and sentence material were found ($p=0.475$). This test shows that listeners could more easily identify words spoken in the CLR style. Moreover, identification was easier for material A, as expected. These results are consistent with the results from previous studies (Picheny *et al.*, 1986) and confirmed that the speech corpus reflects the inherent differences between CLR and CNV speech.

D. Intelligibility experiment 1

1. Stimuli and procedures

The search for acoustic features that are relevant to speech intelligibility can be performed in many ways. Two approaches are "top-down" (starting with the entire space of candidate acoustic features and splitting into smaller sets) and "bottom-up" (starting with individual candidate acoustic features and recombining into larger sets). In pilot studies (Kusumoto *et al.*, 2007, experiment 1), the former approach was used to test a prosodic (HYB-EFDN) versus spectral (HYB-SP) hybridization configuration, without success. One hypothesis as to the failure of those hybrids was that spectral features and duration features cannot be separated without severely impacting intelligibility, due to coarticulation. Therefore, the purpose of experiment 1 was to examine the intelligibility of a HYB speech condition for which spectral and duration features were grouped together, namely, HYB-DSP (see Table I), which replaced phoneme duration, spectrum, and phoneme sequence from CNV speech with those features from CLR speech. A second condition, HYB-EFN, replaced energy, F0, and nonspeech features of CNV speech with those from CLR speech. To calculate the DOC, the intelligibility of the CNV and CLR speech was also measured.

Forty-eight sentences from the LD partitions of materials A and B were used, respectively (96 sentences total), processed in one of the four conditions (12 sentences per condition, per material). The experiment was carried out in a 2×4 Latin Square design (two speech material types and four conditions). Each sentence was heard only once by a listener. The order of the sentences was randomized, and kept the same for all of the subjects. Sentences were counterbalanced, i.e., for conditions *A*, *B*, *C*, and *D*, and a block of consecutive listeners *L*, *L*+1, *L*+2, *L*+3, each sentence was heard in condition *A* by listener *L*, in condition *B* by listener *L*+1, in condition *C* by listener *L*+2, and in condition *D* by listener *L*+3. For each listener, a test administrator first measured the listener's SNR50% for the CNV sentences of the SD partitions of materials A and then B. The resulting values were then used during the immediately following intelligibility test by adjusting the energy of the noise waveform to the measured SNR50% level for that material. Material A was presented in a first session, and Material B in a second session, with a short break between the two sessions. Sentences were considered correct if listeners recalled four out of five key words (both for the initial SNR50% test and the main intelligibility test).

Eight listeners aged 23–39 participated in the experiment. Listeners were informed that they were going to hear syntactically correct sentences of two types, semantically correct and semantically anomalous. They were given one written example sentence of each type. They were instructed to repeat each sentence aloud to the best of their ability and to guess when unsure or when they could not assign meaning to a sentence.

2. Results and discussion

Listeners' average SNR50% levels in the CNV condition for materials A and B were 0.58 and 1.81 dB, respectively (see also Table III). Figure 1(a) shows listeners' raw average intelligibility scores and SDs. In order to determine statistical significance, the raw scores were subjected to the arcsine transformation (Anscombe, 1948) given by the equation

$$x = \arcsin\sqrt{\frac{r+3/8}{n+3/4}}, \tag{2}$$

where r represents the number of sentences a listener recalled correctly, and n represents the number of sentences presented. This transformation was used because the raw intelligibility score differences in percentages are not comparable in a probabilistic sense, especially at the lower and upper ends of the scale (Studebaker, 1985). The intelligibility of CLR speech was significantly ($\alpha=0.05$) increased compared to the intelligibility of CNV speech, using planned, pairwise, two-tailed *t*-tests ($p<0.001$ and $p=0.036$ for materials A and B, respectively). The hybridized condition HYB-DSP yielded a significant increase in intelligibility over the baseline CNV speech ($p=0.003$ and $p<0.001$ for materials A and B, respectively). On the other hand, HYB-EFN did not show a significant difference over CNV levels ($p=0.229$ and $p=0.311$ for materials A and B, respectively).

The DOC was defined as

$$\mathrm{DOC} = \frac{I_{\mathrm{HYB}} - I_{\mathrm{CNV}}}{I_{\mathrm{CLR}} - I_{\mathrm{CNV}}}, \tag{3}$$

where I represents intelligibility levels in percent and the subscript refers to a specific condition. Thus, the DOC of the HYB-DSP condition was 92% and 146% for materials A and B, respectively. The DOC is, of course, subject to listener variation, and DOC values are not precise metrics but are provided to give a sense of how much impact a feature has.

These results indicated that, for this speaker, the combination of phoneme durations, short-term spectrum, and pho-

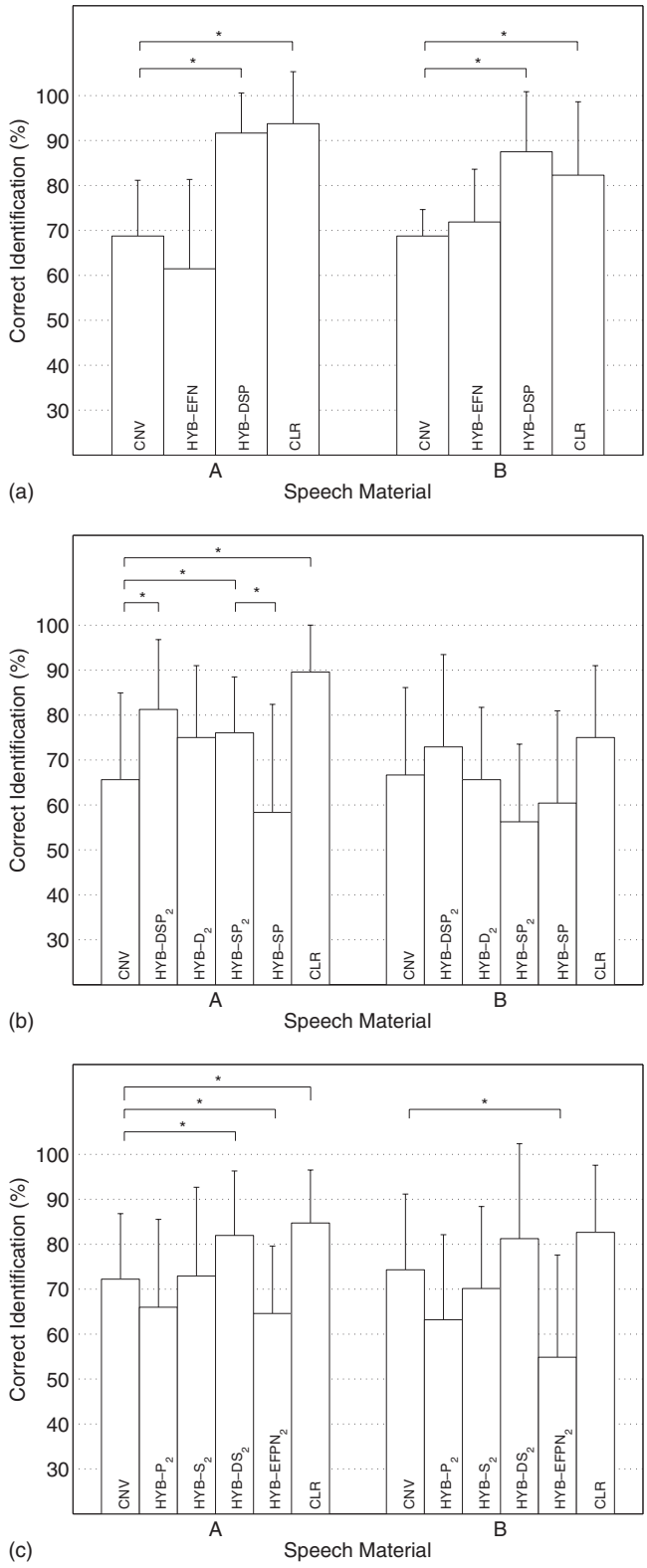

FIG. 1. Raw intelligibility scores for CNV, CLR, and HYB conditions. The error bars indicate the extent of the SDs. The results marked with an asterisk are significantly different (α=0.05).

neme sequence from CLR speech carried most of the information contributing to intelligibility. The combination of prosodic features F0, energy, and nonspeech events (e.g., pauses) did not appear to contribute to intelligibility. However, it was not possible to determine whether duration and spectrum must both be present in the HYB feature set, because the two features were not tested independently; this was further addressed in experiment 2.

E. Improving the algorithm implementation

To further improve speech modification quality, sources of signal processing artifacts were identified and addressed through changes to the hybridization algorithm, leading to a newer implementation. Implementation 2 addressed the following four possible sources of signal-processing artifacts that were hypothesized to impact speech intelligibility and quality.

(1) *Preserving glottalization*. Normally, auxiliary marks were added to the collection of time marks whenever the distance between pitch marks was larger than a minimum threshold (see Sec. III B 3). However, voiced regions may at times have even lower fundamental frequencies than the minimum threshold if the speech sound is glottalized, a phenomenon that occurs when the vocal folds vibrate irregularly or with very low frequency. Therefore, phoneme identity was used to force voiced phonemes to be excluded from having auxiliary marks placed within their region, allowing for very low fundamental frequencies. Specifically, vowels, nasals, flaps, and approximants were excluded from having auxiliary marks added. As a result, the hybridization algorithm could preserve glottalization that was present in an original waveform.

(2) *Preventing artifacts during bursts*. During synthesis, modifications to duration or F0 (when combining spectral information from one style with duration and/or F0 information from another style) may require the duplication of frames. Duplicating frames that contain bursts can cause artifacts because short, single, impulselike acoustic events may sound unnatural when perceived more than once in short succession. Therefore, frames of affricates, plosives, and flaps were prevented from being candidates for duplication; instead, neighboring nonplosive frames were used for achieving the desired duration or F0. Effectively, no duration modification was performed for the phonemes listed above. While this makes the hybridization somewhat less precise, it was hypothesized that the improved perceptual quality would lead to improved intelligibility.

(3) *Preventing artifacts during voicing transitions*. Duplicating frames in unvoiced-to-voiced transitions or in voiced-to-unvoiced transitions may cause artifacts, because these transitions contain unique events that, when duplicated, may be perceived as unnatural. More precisely, these transitory frames are special in terms of their energy, period length, and formant frequency dynamics. Therefore, duplication in frames near unvoiced-to-voiced transitions or voiced-to-unvoiced transitions was prevented; instead, neighboring frames from more steady-state regions were used for processing, similar to the previously described approach to preventing artifacts during bursts.

(4) *Smoothing phoneme insertions and deletions*. Phoneme insertions and deletions, as required by the waveform parallelization stage (see Sec. III B 2), may cause signal discontinuities at phoneme boundaries, which may result in audible clicks, possibly reducing intelligibility. Therefore, during phoneme insertion and deletion operations, all required waveforms were faded in and out smoothly (cross-faded), using linearly weighted windows with durations of approximately 1 ms, centered at the concatenation points.

In summary, implementation 2 allows the occurrence of very low F0 due to glottalization in voiced sounds, prevents frame duplication near voicing transitions or during plosive speech regions such as bursts, and smoothly fades in and out of phoneme insertions and deletions.

F. Intelligibility experiment 2

1. Stimuli and procedures

As observed in experiment 1, the combination of phoneme duration, spectrum, and phoneme sequence of CLR speech significantly improved intelligibility for this speaker, relative to CNV speech. To further test the hypothesis that phoneme duration cannot be separated from spectral features while maintaining high intelligibility (see Sec. IV D), this second experiment used phoneme duration only; spectral features and phoneme sequence; and duration, spectral, and phoneme sequence features from CLR speech. To test the hypothesis that artifacts introduced by the hybridization algorithm can cause decreased intelligibility, the intelligibility of HYB speech produced by implementation 2 was compared with HYB speech that was produced by implementation 1.

Six stimulus conditions were tested in experiment 2. The conditions were CNV speech, CLR speech, and four HYB speech conditions, namely, HYB-DSP_2, HYB-D_2, HYB-SP_2, and HYB-SP, where the subscript "2" indicates hybridization algorithm implementation 2, and the lack of a subscript indicates implementation 1. The HYB-DSP_2 condition examined the combined effects of phoneme duration, spectral features, and phoneme sequences from CLR speech. The HYB-D_2 condition examined the effect of using only phoneme durations from CLR speech. The HYB-SP_2 condition examined the combination of spectral features and phoneme sequences from CLR speech. Finally, the HYB-SP condition was the same in all respects as condition HYB-SP_2, except that it was generated using implementation 1.

As in the previous experiment, a total of 96 sentences from the LD partitions of materials A and B were tested, but this time processed in each of the six conditions described above. The experiment was carried out in a 2×6 Latin Square design (two speech material types and six conditions). Procedures, including SNR50% testing, were identical to experiment 1 (see Sec. IV D 1), except that the number of conditions was six (eight sentences per condition, per material). Twelve listeners aged 19–39 participated.

2. Results and discussion

Listeners' average SNR50% levels in the CNV condition for materials A and B were −0.24 and −0.07 dB, respectively (see Table III). Figure 1(b) shows average intelligibility levels and SDs. Results were transformed using the arcsine transformation [Eq. (2)] and analyzed for statistical significance using a *t*-test, in the same manner as experiment 1. For material A, the intelligibility of CLR speech was significantly higher than the intelligibility of CNV speech ($p<0.001$). However, in material B, a significant difference between CNV and CLR speech was not shown ($p=0.277$). Therefore, the lack of a significant increase between CNV and any hybrid condition for material B was an expected result, as the CLR speech intelligibility level represents the maximum expected performance of HYB speech intelligibility, with the exception of possible listener variation. Planned tests that did not have significant results were not further analyzed. It was speculated that the lack of a significant intelligibility increase of CLR speech may be due to the increased comprehension difficulty of this text material, test-retest reliability, the current group of listeners, or increased variance in response to semantically anomalous sentences.

For material A, a pairwise comparison between HYB-DSP_2 and CNV ($p=0.010$), and between HYB-SP_2 and CNV ($p=0.045$) showed a significant improvement in intelligibility. A comparison between HYB-D_2 and CNV ($p=0.139$) did not show a significant difference. A comparison between HYB-SP_2 and HYB-SP ($p=0.007$) indicated a significant difference in intelligibility favoring implementation 2. The DOCs of the HYB-DSP_2 and HYB-SP_2 conditions were 72% and 56%, respectively.

It appears that, for this speaker, HYB speech that combines spectral, duration, and phonetic sequence features from CLR speech had greater intelligibility than HYB speech from either spectral and phonetic sequence or merely duration features; however, significance testing was not planned for these comparisons. The combination of spectral, duration, and phoneme sequence features, as well as the combination of spectral and phoneme sequence features, yielded significant improvements in intelligibility over CNV speech. Moreover, a significant difference in intelligibility between the two implementations of the hybridization algorithm was found.

G. Intelligibility experiment 3

1. Stimuli and procedures

To determine whether phoneme insertions and deletions between CNV and CLR were significant for intelligibility, the following six stimulus conditions were tested: CNV speech, CLR speech, and the four HYB speech conditions HYB-P_2, HYB-S_2, HYB-DS_2, and HYB-$EFPN_2$. The HYB-P_2 condition examined the effects of using CLR phoneme sequences. The HYB-S_2 condition examined the effect of using only spectral features from CLR speech, similar to the output of a spectral VT system (e.g., Kain and Macon, 1998) with perfect mapping. The HYB-DS_2 condition examined the combined effects of phoneme durations and spectral features from CLR speech. Finally, the HYB-$EFPN_2$ condi-

tion examined the combined effects of energy, F0, phoneme identity, and nonspeech events from CLR speech; this is equivalent to starting out with CLR speech and replacing spectral features and phoneme durations with those from CNV. Experimental conditions were identical to experiment 2, with 18 listeners aged 18–35 participating.

2. Results and discussion

Listeners' average SNR50% levels in the CNV condition for materials A and B were 0.22 and 0.30 dB, respectively (see Table III). Figure 1(c) shows average intelligibility levels and SDs. Results were transformed using the arcsine transformation [Eq. (2)] and analyzed for statistical significance using a *t*-test, in the same manner as previous experiments. For material A, the intelligibility of CLR speech was significantly higher than the intelligibility of CNV speech ($p<0.003$). Similar to experiment 2, material B did not show a significant difference between CNV and CLR speech ($p=0.175$). For material A, comparisons between CNV and hybrid conditions HYB-P_2 and HYB-S_2 showed no significant differences. A comparison between CNV and HYB-DS_2 showed a significant increase in intelligibility ($p=0.050$); the DOC was 78%. In contrast, a comparison between CNV and HYB-$EFPN_2$ showed a significant decrease in intelligibility ($p=0.044$) for both materials.

These results suggest that, for this speaker, the higher intelligibility of CLR speech for material A is not due to either the CLR phoneme sequence or CLR spectrum alone, or due to the combined effects of energy, F0, phoneme sequence, and nonspeech events including pauses. However, the combined effects of spectral features and phoneme durations yielded a significant intelligibility increase compared to CNV speech.

V. CONCLUSION AND FUTURE WORK

This article presented experimental results of sentence-level improvements to the intelligibility of CNV speech, by using hybrid stimuli which combine certain features of a CNV speech sentence with complementary features of the same CLR speech sentence. The condition HYB-DSP, for example, is equivalent to a LPC resynthesis of CLR speech with F0 and energy trajectories modified to match F0 and energy trajectories of CNV speech; additionally, nonspeech parts (pauses) were removed, according to the labels of the CNV speech of that same sentence. Listening tests estimated the significance and DOC of acoustic features, using normal-hearing listeners aged 19–41 and 12-speaker babble noise.

There are two main conclusions from this work. First, it is possible to create HYB speech that combines aspects of CNV and CLR speech to create a signal with greater intelligibility than the original CNV speech. Second, the results indicate that the two main sources of increased intelligibility of CLR speech for this speaker and material A are in the spectrum and duration, and not in the pausing patterns, F0, energy, or phoneme sequence. In particular, for material A, experiment 1 showed that the feature combination DSP was significant for improved intelligibility, while the combination EFN was not. Experiment 2 then showed that while the combination DSP was significant, and SP alone provided a significant contribution, duration alone was not considered significant. Experiment 3 then showed that DS alone was significant for improved intelligibility, but that neither spectrum nor phoneme sequence were significant by themselves. (Future experiments will be required to better assess the interactions between spectrum, duration, and phoneme sequence.) These results present the first known study in which sentence-level intelligibility of CNV speech has been improved by application of a subset of CLR speech features.

We speculate that the spectrum is important for intelligibility because of the general differences in the size of the vowel space between CLR and CNV speech (Picheny *et al.*, 1986). In addition, spectral tilt (included in our definition of spectrum) may also play a role in intelligibility. Duration, and in particular the combination of duration and spectrum, may be important to intelligibility because of formant dynamics. If formants are stretched or compressed in a way that is unnatural, the unusual formant dynamics may negatively impact intelligibility. Also, relative duration may affect phoneme characteristics such as the voicing of stops or the tense/lax quality of a vowel, and so control of duration may be important when matched with acoustic cues such as periodicity or formant locations.

Differences in intelligibility scores were more pronounced for material A than for material B. It was conjectured that the difference in intelligibility levels between materials may be caused by easier comprehension of material A, characterized by CNV and CLR psychometric functions with steeper slopes, leading to larger differences at or close to SNR50%. A possible increase in the variability of listener responses to semantically anomalous sentences may also be a contributing factor. The lack of significant differences between CNV and CLR for material B in experiments 2 and 3 prevented comparisons between CNV and HYB conditions in those experiments.

It is important to note that these results are for one speaker only and cannot be generalized to the larger population, since speakers use different strategies for speaking clearly (Ferguson and Kewley-Port, 2002; Ferguson, 2004; Krause and Braida, 2004). In addition, results may be specific to the particular corpus used.

The hybridization algorithm is equivalent to modifying CNV speech with an "oracle" mapping function, thus simulating maximum performance levels of an automatic modification system. An automatic modification system with a trained mapping function has potential to improve speech intelligibility for hearing-impaired listeners, such as individuals with hearing loss or in noisy environments. Speech modification of duration may be undesirable in live environments where speakers are both audible and visible, because of the resulting asynchrony between lip movements and auditory events. However, duration modification may be useful in live audio-only settings such as telephone-based speech. In such applications, long-term asynchrony might be compensated for by either additional feedback requesting the speaker to pause when necessary, or automatically modifying the length of pauses to compensate for changed speech durations. Moreover, recorded or bufferable content (such as

hard-disk-based viewing of TV shows) could be enhanced by modifying the video portion of the signal as well, preserving perceptually relevant synchronisms.

Future work will use the current methods for determining the importance of features such as formant dynamics, spectral tilt, and relative duration, using several speakers and several listener types (different age groups and hearing performances). The size of the units studied may also be varied. For example, spectral changes were applied at the phoneme level in the current study. Larger spectral units [e.g., syllable-level features, used in voice transformation (Rao *et al.*, 2007)] or smaller units (e.g., phoneme-transition regions) may also yield information of interest. This work may lead to objective measures of speech intelligibility and automatic speech intelligibility enhancement systems. The automatic mapping of CNV speech features to more closely resemble CLR speech features will not be a research focus until it is better known which specific acoustic features are relevant for the improved speech intelligibility of CLR speech.

ACKNOWLEDGMENTS

The authors thank Jan P. H. van Santen of OHSU, Marjorie R. Leek and Michelle Molis of the National Center for Rehabilitative Auditory Research at the Portland Veterans Affairs Medical Center for insightful discussions and advice, as well as the reviewers for their helpful comments and suggestions.

Abe, M. (**1996**). "Speech morphing by gradually changing spectrum parameter and fundamental frequency," in Proceedings of ICSLP, Philadelphia, PA, Vol. **4**, pp. 2235–2238.

Anscombe, F. J. (**1948**). "The transformation of poisson, binomial and negative-binomial data," Biometrika **35**, 246–254.

Bilger, R. C., and Nuetzel, J. M. (**1984**). "Standardization of a test of speech perception in noise," J. Speech Hear. Res. **27**, 32–48.

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," Proc. Inst. Phonetic Sci. **17**, 97–110.

Bond, Z. S., and Moore, T. J. (**1994**). "A note on the acoustic-phonetic characteristics of inadvertently clear speech," Speech Commun. **14**, 325–337.

Bradlow, A. R., and Bent, T. (**2002**). "The clear speech effect for non-native listeners," J. Acoust. Soc. Am. **112**, 272–284.

Bradlow, A. R., Krause, N., and Hayes, E. (**2003**). "Speaking clearly for children with learning disabilities: Sentence perception in noise," J. Speech Lang. Hear. Res. **46**, 80–97.

Bradlow, A. R., Torretta, B. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun. **20**, 255–272.

Center for Spoken Language Understanding (**2007**). "Clear Speech JPH," http://www.cslu.ogi.edu/corpora (last viewed April, 2008).

Chang, J. E., Bai, J. Y., and Zeng, F.-G. (**2006**). "Unintelligible Low-Frequency Sound Enhances Simulated Cochlear-Implant Speech Recognition in Noise," IEEE Trans. Biomed. Eng. **53**, 2598–2601.

Cruickshanks, K., Wiley, T., Tweed, B., Klein, B., Klein, R., Mares-Perlman, J., and Nondahl, D. (**1998**). "The prevalence of hearing loss in older adults," Am. J. Epidemiol. **148**, 879–885.

Dillon, H. (**2001**). *Hearing Aids* (Thieme, New York).

Ferguson, S. H. (**2004**). "Talker differences in clear and conversatoinal speech: Vowel intelligibility for normal-hearing listeners," J. Acoust. Soc. Am. **116**, 2365–2373.

Ferguson, S. H., and Kewley-Port, D. (**2002**). "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **112**, 259–271.

Gordon-Salant, S. (**1986**). "Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing," J. Acoust. Soc. Am. **82**, 1599–1607.

Gordon-Salant, S. (**1987**). "Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects," J. Acoust. Soc. Am. **81**, 1199–1202.

Gordon-Salant, S., and Fitzgibbons, P. J. (**1997**). "Selected cognitive factors and speech recognition performance among young and elderly listeners," J. Speech Lang. Hear. Res. **40**, 423–431.

Gordon-Salant, S., and Fitzgibbons, P. J. (**2001**). "Sources of age-related recognition difficulty for time-compressed speech," J. Speech Lang. Hear. Res. **44**, 709–719.

Hazan, V., and Markham, D. (**2004**). "Acoustic-phonetic correlates of talker intelligibility for adults and children," J. Am. Acad. Audiol **116**, 3108–3118.

Hazan, V., and Simpson, A. (**1998**). "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," Speech Commun. **24**, 211–226.

Helfer, K. S. (**1998**). "Auditory and auditory-visual recognition of clear and conversational speech by older adults," J. Am. Acad. Audiol **9**, 234–242.

Hosom, J. (**2002**). "Automatic phoneme alignment based on acoustic-phonetic modeling," in Proceedings of ICSLP, Boulder, CO, Vol. **1**, pp. 357–360.

International Electrotechnical Commission (**2002**). "Electroacoustics-sound level meters—Part 1: Specifications," Paper No. 61672.

Kain, A. (**2001**). "High resolution voice transformation," Ph.D. thesis, Oregon Graduate Institute, Portland, OR.

Kain, A., Hosom, J.-P., Niu, X., van Santen, J., Fried-Oken, M., and Staehely, J. (**2007**). "Improving the intelligibility of dysarthric speech," Speech Commun. **49**, 743–759.

Kain, A., and Macon, M. (**1998**). "Spectral voice conversion for text-to-speech synthesis," in Proceedings of ICASSP, Vol. **1**, pp. 285–288.

Krause, J. C., and Braida, L. D. (**2002**). "Investigation alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility," J. Acoust. Soc. Am. **112**, 2165–2172.

Krause, J. C., and Braida, L. D. (**2004**). "Acoustic properties of naturally produced clear speech at normal speaking rates," J. Acoust. Soc. Am. **15**, 362–378.

Kusumoto, A., Kain, A., Hosom, J.-P., and van Santen, J. (**2007**). "Hybridizing conversational and clear speech," in Proceedings of Interspeech.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Liu, S., Rio, E. D., Bradlow, A. R., and Zeng, F. G. (**2004**). "Clear speech perception in acoustic and electric hearing," J. Acoust. Soc. Am. **116**, 2374–2383.

Liu, S., and Zeng, F. G. (**2006**). "Temporal properties in clear speech perception," J. Acoust. Soc. Am. **120**, 424–432.

Nejime, Y., and Moore, B. C. J. (**1998**). "Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss," J. Acoust. Soc. Am. **103**, 572–576.

Payton, K. L., Uchanski, R. M., and Braida, L. D. (**1994**). "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," J. Acoust. Soc. Am. **95**, 1581–1592.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1985**). "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," J. Speech Hear. Res. **28**, 96–103.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," J. Speech Hear. Res. **29**, 434–446.

Pichora-Fuller, M. K., Schneider, B. A., and Daneman, M. (**1995**). "How young and old adults listen to and remember speech in noise," J. Acoust. Soc. Am. **97**, 593–608.

Rabiner, L., and Juang, B. H. (**1993**). *Fundamental of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).

Rao, K. S., Laskar, R. H., and Koolagudi, S. G. (**2007**). "Voice transformation by mapping the features at syllable level," in Pattern Recognition and Machine Intelligence, Heidelberg, Germany, Vol. **4815**, pp. 479–486.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silberger, H. R., Urbanek, G. E., and Weinstock, M. (**1969**). "IEEE Recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 227–246.

Salthouse, T. A. (**1996**). "The processing-speech theory of adult age differences in cognition," Psychol. Rev. **103**, 403–428.

Schum, D. J. (**1996**). "Intelligibility of clear and conversational speech of young and elderly talkers," J. Am. Acad. Audiol **7**, 212–218.

Studebaker, G. A. (**1985**). "A "rationalized" arcsine transform," J. Speech Hear. Res. **28**, 455–462.

Taylor, P., Black, A., and Caley, R. (**1998**). "The architecture of the festival speech synthesis system," in Proceedings of the Third International Workshop on Speech Synthesis, Sydney, Australia.
Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., and Durlach, N. I. (**1996**). "Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate," J. Speech Hear. Res. **39**, 494–509.
Vaughan, N., Storzbach, D., and Furukawa, I. (**2006**). "Sequencing and nonsequencing working memory in understanding of rapid speech by older listeners," J. Am. Acad. Audiol **17**, 506–518.
Wingfield, A., Tun, P. A., and McCoy, S. L. (**2005**). "Hearing loss in older adulthood. What it is and how it is interacts with cognitive performance," Current directions in Psychological Science **14**, 144–148.