Published in final edited form as:

Speech Commun. 2014 April 1; 59: 1–9. doi:10.1016/j.specom.2013.12.001.

Determining the relevance of different aspects of formant contours to intelligibility[☆]

Akiko Amano-Kusumoto^a, John-Paul Hosom^{b,1}, Alexander Kain^b, and Justin M. Aronoff^{a,2}

^aDepartment of Human Communication Science Devices, House Research Institute 2100 West Third Street, Los Angeles, California 90057

^bCenter for Spoken Language Understanding (CSLU), Oregon Health & Science University (OHSU) 20000 NW Walker Road, Beaverton, Oregon 97006

Abstract

Previous studies have shown that “clear” speech, where the speaker intentionally tries to enunciate, has better intelligibility than “conversational” speech, which is produced in regular conversation. However, conversational and clear speech vary along a number of acoustic dimensions and it is unclear what aspects of clear speech lead to better intelligibility. Previously, Kain et al. [J. Acoust. Soc. Am. **124** (4), 2308–2319 (2008)] showed that a combination of short-term spectra and duration was responsible for the improved intelligibility of one speaker. This study investigates subsets of specific features of short-term spectra including temporal aspects. Similar to Kain’s study, hybrid stimuli were synthesized with a combination of features from clear speech and complementary features from conversational speech to determine which acoustic features cause the improved intelligibility of clear speech. Our results indicate that, although steady-state formant values of tense vowels contributed to the intelligibility of clear speech, neither the steady-state portion nor the formant transition was sufficient to yield comparable intelligibility to that of clear speech. In contrast, when the entire formant contour of conversational speech including the phoneme duration was replaced by that of clear speech, intelligibility was comparable to that of clear speech. It indicated that the combination of formant contour and duration information was relevant to the improved intelligibility of clear speech. The study provides a better understanding of the relevance of different aspects of formant contours to the improved intelligibility of clear speech.

Keywords

speech intelligibility; vowel perception; speech synthesis

[☆]Part of this study was presented in [1].

© 2013 Elsevier B.V. All rights reserved.

Correspondence to: Akiko Amano-Kusumoto.

¹Currently at Sensory Inc.

²Currently at Department of Speech and Hearing Science, University of Illinois at Urbana-Champaign.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

People often adopt a particular speaking style to better communicate with hard-of-hearing or cognitively-impaired individuals, referred to as *clear* speech. Previous research has found that clear speech is more intelligible than *conversational* speech, which is spoken in regular communication [2]. The acoustic-phonetic characteristics of clear speech have been shown to differ from those of conversational speech [3]. For example, clear speech has slower speaking rate, expanded vowel space, increased energy at higher frequency regions, and lengthened phoneme durations [3, 4, 5]³. Given the various acoustic-phonetic characteristics of clear speech, what makes clear speech more intelligible is still questionable.

To address the question, Kusumoto et al. [7] and Kain et al. [8] examined the contribution of certain acoustic features to improved intelligibility of clear speech by using a *hybridization* algorithm. The hybridization algorithm is a tool for creating hybrid stimuli that have certain features of clear speech utterance, and complementary features of conversational speech. The hybrid speech makes it possible to examine the perceptual relevance of certain features to the improved intelligibility of clear speech. Perceptual experiments indicated that hybrid speech with short-term spectra and phoneme durations modeled after clear speech yielded significant improvement in intelligibility over unprocessed conversational speech [8]. Thus, the authors concluded that the combination of short-term spectra and phoneme durations caused the improved intelligibility of clear speech. However, the short-term spectrum contains a large number of features, therefore, the goal of this study was to investigate which specific features of short-term spectra, including the temporal aspect, can yield better intelligibility. The outcome of this study may be useful for assistive listening devices, which could provide speech with increased intelligibility for listeners with hearing impairment or in adverse listening conditions by enhancing relevant features.

Formant frequencies are the spectral peaks of the short-term spectrum, resulting from the positions of the articulators (e. g. tongue and lip positions) in the oral cavity. The formant contours, the time course of formant frequencies, reflect the movement of articulators from one position to another over time. Previous studies showed that the formant contour over the course of a vowel is perceptually important [9] and that word intelligibility is significantly correlated with mean word duration and the difference between the F2 values of /i/ and /u/ [10]. Therefore, in the first hybrid condition (_{HYB-C}), we examined whether the formant contour was relevant to improved intelligibility. We excluded other features from the short-term spectra such as formants over 4 kHz, formant residuals, and formant bandwidths that were not reported to be different between conversational and clear speech in most cases [5].

However, the entire formant contour may not be responsible for the improved intelligibility of clear speech, and a subset of the formant contour may be sufficient. Furui [11] showed that the formant frequencies at the maximum spectral transition are sufficient to identify vowels. Additionally, Moon and Lindblom [12] showed that the steady-states at the middle point of the vowel are shown to be different between conversational and clear speech, and that F2 at the transition moves faster in clear than in conversational speech. In the second

³For a more complete review, see [6].

hybrid condition ($_{\text{HYB-MT}}$), we examine the relevancy of formant frequencies at the middle point and formant transition, and removing other features (points between transition and middle point, and duration of clear speech) from the first condition, $_{\text{HYB-C}}$.

Finally, it may be sufficient to have formant frequencies from clear speech at the middle point of the vowel but not the formant transition of clear speech. Previous studies showed that the vowel space representing F1 and F2 values was expanded along both dimensions in the clear speaking style [13, 5], and that formant steady-states at the middle point of the vowel in conversational speech do not reach their targets as much compared to clear speech ($_{\text{wVl}}$ context) [12]. The extreme steady-state values of clear speech may be the cause of improved intelligibility. Therefore, the third hybrid condition ($_{\text{HYB-M}}$) was used to examine the relevancy of the formant values at the middle point of the vowel (defined as steady-state values). In summary, this study examined the relative contribution of formant contour, a combination of steady-state formant values and formant transition, or steady-state values alone, and whether these features resulted in an improved intelligibility compared to conversational speech.

2. Speech Corpus

Speech materials containing conversational and clear speech were recorded for this study, using a previously developed word list [12]. This list contains $_{\text{wVl}}$ words designed to achieve (1) large consonant-vowel formant transitions, (2) the same degree of stress on the test word, and (3) systematic change of the vowel duration. The four front vowels ($/i/$, $/ɪ/$, $/ε/$ and $/ei/$) surrounded by the consonants $/w/$ and $/l/$ were recorded.

Four words (*wheel*, *will*, *well*, and *whale*) in a carrier sentence were repeated 16 times each. The carrier sentence “it’s easy to tell the size of a *WORD*” was used to facilitate the use of fundamental frequency (F0), duration and intensity as necessary, upon the elicitation of conversational and clear speech. A total of 128 tokens were recorded ($4 \text{ words} \times 2 \text{ speaking styles} \times 16 \text{ repetitions}$).

The speech signals were recorded digitally at a sampling rate of 16 kHz with 16-bit resolution. One male, a native speaker of North-American English with no professional training in public speaking, was recruited as a speaker because of his previous experience producing conversational and clear speech in the laboratory [8]. When recording conversational speech, the speaker was instructed to recite the materials in a way that he would use to communicate in his daily life. When recording clear speech, he was instructed to speak clearly, as he would talk when communicating with elderly listeners or hearing-impaired listeners. The recording of conversational speech was followed by clear speech in the first and second recording sessions, respectively, using the speaker’s own distinction between conversational and clear speech production.

The average speaking rates in words per minute (wpm), measured excluding pause durations, were 366 wpm and 179 wpm for conversational ($_{\text{CNV}}$) and clear ($_{\text{CLR}}$) speech, respectively. The results of the t -test showed the speaking rate was significantly different between conversational and clear speech ($p < 0.001$).

3. Hybridization algorithm

To examine the effect of a specific feature (or a combination of features) independent of other features, a hybridization algorithm was employed. The hybridization algorithm is a tool for creating hybrid stimuli that have certain features of clear speech utterances, and complementary features of conversational speech. The algorithm involves modifying features of conversational speech to make them more like clear speech. If certain hybrid stimuli had better speech intelligibility than the baseline conversational speech, it would suggest that the clear speech features in the hybrid stimuli play a role in the improved intelligibility of clear speech.

The hybridization algorithm began with the analysis of acoustic features. First, all test words were annotated and segmented automatically using forced alignment [14]. Formant contours of the word of interest and glottal closure instants (GCIs) were extracted using the Snack Sound Toolkit [15]. GCIs were required to analyze waveforms pitch-synchronously. The inverse of one pitch period corresponds to the F0. A trained transcriber manually corrected phoneme boundaries, formant contours and GCIs.

Steady-state formant frequencies, formant transition, and phoneme durations were measured as follows. The steady-state values for F1 through F4 were extracted at the midpoints of each phoneme /w/, /V/, and /l/ of conversational and clear speech, and averaged over the 16 repetitions available per word. Formant slopes for F1 through F4 were measured at the vowel onset and offset centered on the boundary by fitting a straight line to the formant values (three data points) over a 20 ms window. Phoneme durations were measured from the onset and offset of each phoneme.

The next step was to create the hybrid formant contours with a combination of conversational and clear speech (Sections 3.1). Then, the formant values of conversational speech were modified to match the hybrid formant contours, and the hybrid stimuli were synthesized (Section 3.2).

3.1. Creating hybrid formant contours

Three hybrid conditions were evaluated to test the relevance of different aspects of formant contours for intelligibility. The formant contour shape of each condition is shown in Figure 2.

(a) HYB-C: Clear formant contours—This condition is considered the closest to clear speech, examining the formant frequency contour (not the residuals of the formant, higher formant above 4 kHz or formant bandwidth) as well as phoneme duration. Phoneme durations of conversational speech were stretched to match clear speech at the synthesis stage. Therefore, formant contours from clear speech were copied as _{HYB-C} formant contours (see Figure 2).

(b) HYB-MT: Clear steady-state formant frequencies at midpoints and formant transitions—This condition is one step closer to conversational speech from the _{HYB-C} condition. The process of creating a _{HYB-MT} formant contour required a weighting function for

each formant contour (F1 through F4) (see Figure 1). In designing the weighting function, the ratio was calculated between clear and conversational formant values (F1 through F4) at midpoints of the phonemes, and three points at the phoneme boundaries (one on the boundary, and one on each side of the boundary) for both /w/ to /V/ and /V/ to /l/ (see Figure 2). The points at the beginning and ending of the weighting function were set to a ratio of 1.0 to avoid any discontinuities from the unmodified preceding and following waveform. The weighting function for each formant was designed to be linearly ascending or descending to describe the desired ratio. Then, the original conversational formant contour was multiplied by the weighting function. In this way, the formant contours of _{HYB-MT} were set to have the steady-state values at midpoint and transitions of clear speech.

(c) HYB-M: Clear steady-state formant frequencies at midpoints—_{HYB-M} contour reaches clear steady-state values, while duration, residuals of the formant, higher formants (above 4 kHz), and formant bandwidths remain the same as conversational speech. This condition has the fewest clear speech features. The weighting function was designed to be linearly ascending or descending to describe the ratio between the clear steady-state values and the conversational steady-state values at the midpoint of each phoneme, using phoneme duration from conversational speech (see Figure 1). Then, the original conversational formant contour was multiplied by the weighting function to obtain the _{HYB-M} formant contours (see Figure 2).

3.2. Modifying formant values

The process of modifying formant values to match the above described hybrid formant contours involves waveform analysis, removal of existing formant values, and the application of newly designed formant values. During the synthesis of hybrid speech, the pitch-synchronous residual-excited overlap-add method was used.

First, the speech waveform was analyzed with a pitch-synchronous frame that spans two pitch periods with a one pitch period overlap. The speech signal $S(z)$ (z -transformation) can be represented as $S(z) = Q(z) \cdot V(z)$, where $Q(z)$ was the residual signal, and $V(z)$ was a formant filter. $V(z)$ was modeled with complex pole pairs represented as

$$V(z) = \frac{1}{\prod_{k=1}^4 (1 - r_k e^{j\theta_k} z^{-1})(1 - r_k e^{-j\theta_k} z^{-1})} \quad r_k = e^{-\pi b_k / F_s}, \theta_k = \frac{2\pi f_k}{F_s} \quad (1)$$

where f_k was the k -th formant frequency, b_k was its bandwidth, and F_s was the sampling frequency.

The spectral envelope at one frame was then removed by applying an inverse filter of the formant filter, which was designed with formant frequencies of conversational speech ($Q_{CNV}(z) = S_{CNV}(z)/V_{CNV}(z)$). The residual signal contained primarily the glottal source and higher formants of conversational speech. _{HYB-C} required stretching the phoneme duration to match that of clear speech. In this condition, after the inverse filtering, a frame of residual signal was repeated as necessary to obtain the desired clear phoneme durations by linear sampling throughout the duration of the phoneme.

The hybrid formant contours ($_{HYB-C}$, $_{HYB-MT}$, and $_{HYB-M}$) were used to design time-varying all-pole digital filters ($V_{HYB}(z)$). The bandwidths of each filter were unchanged from those of the original conversational speech. The speech waveform at each frame was obtained by applying the all-pole digital filters to the residual signal ($S_{HYB}(z) = V_{HYB}(z) \cdot Q_{CNV}(z)$). For the overlap-add operation, formant-filtered waveform was windowed with an asymmetric trapezoidal shape window [16]. The second half of one frame was then added to the first half of the next frame.

4. Stimuli

Four vowels (/i/, /ɪ/, /ε/ and /ei/) in /wV/ context with 16 repetitions were tested in five conditions ($_{CLR}$, $_{HYB-C}$, $_{HYB-MT}$, $_{HYB-M}$, and $_{CNV}$). Isolated words, without the carrier sentence, were presented to the subjects.

In order to examine the effect of formant contour, the F0 contour of the test word and energy of the vowel were normalized across all speech samples. First, an F0 contour template was derived from naturally fast spoken clear speech from the pilot study based on the F0 onset values of each phoneme (/w/, /V/, /l/), the F0 offset values of the /l/, and the maximum F0 value in /V/. The five points in the F0 contour are shown in Table 1. The F0 contour template was then derived by interpolating these five points with cubic spline interpolation, and stretched (or compressed) to match the duration of each phoneme with the observed data in each speaking condition. The fast spoken clear speech was chosen for the F0 contour template because that obviated the need of lowering F0 values for the conversational speech, potentially resulting in less audible signal artifacts. In the F0 modification stage, the F0 values of each sample were modified to the values of the F0 contour template. The length of time frame that spans two pitch-period was changed to match the new pitch-period based on the F0 contour template. In the synthesis stage, the pitch-synchronous residual-excited overlap-add method was used to synthesize the final waveform.

The energy of the vowel was normalized across all speech samples. First, the A-weighted root mean square value of the vowel in a test word ($RMS_v(w)$) was calculated [17]. The gain factor was then obtained by dividing the constant RMS_v by $RMS_v(w)$ for a particular word. Finally, the waveform of the test word was multiplied by the gain factor, which resulted in constant RMS_v for the vowel in all test words.

5. Procedure and apparatus

The experiment took place in a quiet room, where a listener was seated in front of a computer monitor, listening to stimuli through circumaural headphones (Sennheiser HD 280 Pro), binaurally. 12-talker babble noise was added to bring the intelligibility level near a threshold which is sensitive to changes in performance across conditions. The level of the noise was adjusted to the signal-to-noise ratio (SNR) at which the listener could correctly identify $_{CNV}$ stimuli 50 % of the time (SNR=50) combining all four words pooled together. The speech level was set to listeners' most comfortable level. After the practice, the SNR=50 level was obtained for each listener using an adaptive procedure to minimize the between-subject variability. The rest of the tests were conducted at the fixed SNR level (SNR=50). For both adaptive and fixed level procedures, the listener's task was a forced-

choice test, corresponding to the four choices (“wheel”, “will”, “well”, “whale”). The order of 320 stimuli (4 /wV/ words $\times$ 5 conditions $\times$ 16 repetitions) was randomized once, and kept the same for all of the listeners while rotating conditions.

6. Subjects

Fifteen adults (average age 32.9 years old, standard deviation 11.6: 8 males and 7 females) participated in the perceptual experiment. All were native speakers of North-American English with self-reported normal hearing.

7. Results

The percent correct rates averaged over 15 listeners and four vowels were examined to determine whether the intelligibility was significantly improved with clear speech (see Figure 3). The effects of condition ($_{\text{CNV}}$ and $_{\text{CLR}}$) and vowels (4 vowels) were tested using a two-way repeated measures analysis of variance (ANOVA). The main effect of condition was significant ($F(1, 112) = 53.51, p < 0.001$). While there was no significant main effect of vowel ($p > 0.05$), there was a significant interaction between condition and vowel ($F(3, 112) = 17.89, p < 0.001$). Post-hoc pairwise comparisons were conducted between conversational and clear speech for each vowel. The results showed that the intelligibility of the vowels /i/ and /ei/ significantly improved with clear speech ($p < 0.001$ for both), while the intelligibility of the vowels /ɪ/ and /ɛ/ did not ($p > 0.05$ for both). As a result, further analysis of the hybrid speech focused on the vowels /i/ and /ei/. It is worth noting that the average intelligibility of conversational speech was higher for the vowels /ɪ/ and /ɛ/ than the vowels /i/ and /ei/, even though noise level was set at 50 % for conversational speech for all four words.

The effects of condition ($_{\text{CLR}}$, $_{\text{HYB-C}}$, $_{\text{HYB-MT}}$, $_{\text{HYB-M}}$, and $_{\text{CNV}}$) and vowel (2 vowels) were tested using a two-way repeated measures analysis of variance (ANOVA). The main effect of condition ($F(4, 140) = 55.25, p < 0.001$) and the vowel ($F(1, 140) = 12.33, p < 0.001$) were both significant. There was also a significant interaction between condition and vowel ($F(4, 140) = 9.04, p < 0.001$). Post-hoc pairwise comparisons showed that for the vowel /i/ all three hybrid conditions yielded improved intelligibility over the $_{\text{CNV}}$ speech. Both $_{\text{CLR}}$ and $_{\text{HYB-C}}$ were better than $_{\text{HYB-MT}}$ and $_{\text{HYB-M}}$ ($p < 0.01$ for all), while $_{\text{CLR}}$ and $_{\text{HYB-C}}$ were not statistically different ($p > 0.05$). For the vowel /ei/, two of the hybrid conditions, $_{\text{HYB-C}}$ and $_{\text{HYB-M}}$, were significantly better than $_{\text{CNV}}$ ($p < 0.01$ for both), but $_{\text{HYB-MT}}$ was not ($p > 0.05$). $_{\text{HYB-MT}}$ was worse than the $_{\text{HYB-M}}$ condition ($p < 0.0001$). Similar to the vowel /i/, both $_{\text{CLR}}$ and $_{\text{HYB-C}}$ were better than $_{\text{HYB-MT}}$ and $_{\text{HYB-M}}$ ($p < 0.0001$).

The intelligibility of $_{\text{HYB-MT}}$ and $_{\text{HYB-M}}$ were not statistically different for the vowel /i/, which suggested adding clear speech formant transition did not have much effect on intelligibility. For the vowel /ei/, on the other hand, adding clear speech formant transition had negative impact on intelligibility.

The results showed that for the tense vowel, modifying formant frequencies with the steady-state formant frequency of clear speech has the potential to improve intelligibility of conversational speech, but not to yield intelligibility comparable to that of clear speech.

However, when the entire formant contour of conversational speech was replaced by that of clear speech, intelligibility was comparable to that of clear speech.

8. Discussion

In the experiment, the intelligibility of artificially modified hybrid speech representing intermediary aspects between clear and conversational speech was examined. Three hybrid conditions were tested investigating: (1) the benefit of the entire formant contour of clear speech, (2) the benefit of the steady-state formant frequency and formant transition of clear speech, and (3) the benefit of the steady-state formant frequency of clear speech,

The results showed that replacing the entire formant contour of clear speech ($_{HYB-C}$) was the most effective to improve the intelligibility of conversational speech. It indicated that the combination of formant contour and duration information was relevant to the improved intelligibility of clear speech. Similarly, replacing the steady-state formant frequency ($_{HYB-M}$) was also effective to improve intelligibility, but not to that of the clear speech.

The perceptual results were consistent with the acoustic changes between conversational and clear speech (See Table 2 and Figure 4). For tense vowels where clear speech was more intelligible than conversational speech, the acoustic differences of F2 steady-states and vowel durations were larger than lax vowels. The amount of acoustic change found in lax vowels might not have been sufficient enough to improve the intelligibility of clear speech over that of conversational speech.

The result showed that when the entire formant contour of conversational speech was replaced by that of clear speech, intelligibility was comparable to that of clear speech for the tense vowels. This suggests that the entire formant contour is important for perception, not only steady-state values or formant transitions, as previously suggested [9].

Adding formant transition information with steady-state values did not help, which may be because $_{HYB-M}$ dominated the improvement over the conversational speech. On the other hand, $_{HYB-MT}$ was significantly worse than $_{HYB-M}$ for the vowel /ei/. One possible explanation for this negative results is that the transition, in addition to the steady-state values at the midpoint, disrupted the shape of the formant contour of the diphthong.

The results suggest that the residual signal (glottal source and higher formants) and bandwidths from clear speech may not be relevant to improved intelligibility, although we cannot rule out the possibility that there are other co-occurring changes in clear speech. The hybrid speech was synthesized by applying a formant filter ($V_{HYB}(z)$) to the residual signal ($Q_{CNV}(z)$). Also, the bandwidths of the formant filter ($V_{HYB}(z)$) were unchanged from those of conversational speech. Yet, results showed that hybrid speech had improved intelligibility over conversational speech.

It is important to note that this study involved one speaker and a subset of vowels and further studies are needed to determine how widely these results will generalize. Prior work has found that different speakers might employ different strategies, when they are instructed to produce clear speech [4, 18]. Moreover, the improved intelligibility of clear speech has

been shown with different types of listeners, including normal-hearing listeners [19, 20, 21], elderly normal-hearing listeners [22], elderly hearing-impaired listeners [4], and cochlear-implant users [21]. It might be necessary to test specific features for target populations in the future. For example, elderly hearing-impaired listeners would probably benefit from lengthened duration more than young listeners due to slowed speed of processing [23].

The _{HYB-C} condition includes an inherent timing property. The question of whether clear formant contours with conversational speech duration would yield the intelligibility of clear speech is still unclear. A possible hybrid condition with the formant contours from clear speech and duration from conversational speech could be implemented by compressing clear formant contours in the future. It might be possible to compress the steady-state portions and preserve transition, although the steady-state formant frequency might not reach as far as clear speech steady-state values due to its short duration.

This work is an important step towards understanding what aspects of clear speech lead to better intelligibility. In summary, the formant modification targeting clear steady-state formant frequencies was effective to significantly improve vowel intelligibility as compared with conversational speech. On the other hand, modifying the entire formant contour further improved the intelligibility of conversational speech to a level comparable to that of clear speech. Thus, the entire formant contour is perceptually relevant to the intelligibility of clear speech.

9. Conclusions

In this work, the perceptual relevance of various acoustic feature to the improved intelligibility of clear speech was studied. To manipulate specific features independent of other features, hybrid stimuli were created where portions of conversational speech were selectively replaced by formant features from clear speech. The results showed that although steady-state formant frequencies contributed to the improved intelligibility of clear speech, neither the steady-state portion nor the formant transition was sufficient to yield intelligibility comparable to that of clear speech. Additionally, when the entire formant contour of conversational speech was replaced by that of clear speech, including duration from clear speech, intelligibility was comparable to that of clear speech for the tense vowels. The results suggest that modifying the entire formant contour with duration from clear speech may be an effective way to improve intelligibility for listeners with hearing impairment or in adverse listening conditions, and that the entire formant contour is important for perception, not only steady-state values or formant transitions. A future study will investigate how to capture the formant contour of clear speech while keeping the phoneme duration of conversational speech.

References

1. Amano-Kusumoto A, Hosom J-P. The effect of formant trajectories and phoneme durations on vowel intelligibility. *Proc. of ICASSP*. 2009:4677–4680.
2. Picheny MA, Durlach NI, Braida LD. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*. 1985; 28:96–103. [PubMed: 3982003]

3. Picheny MA, Durlach NI, Braida LD. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*. 1986; 29:434–446. [PubMed: 3795886]
4. Ferguson SH, Kewley-Port D. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*. 2002; 112:259–271. [PubMed: 12141351]
5. Krause JC, Braida LD. Acoustic properties of naturally produced clear speech at normal speaking rates. *Journal of the Acoustical Society of America*. 2004; 115:362–378. [PubMed: 14759028]
6. Amano-Kusumoto A, Hosom J-P. A review of research on speech intelligibility and correlations with acoustic features. CSLU Technical Report. 2011; 11-002:1–16.
7. Kusumoto A, Kain A, Hosom J-P, van Santen J. Hybridizing Conversational and Clear Speech. *Proc. of Interspeech*. 2007:370–373.
8. Kain A, Amano-Kusumoto A, Hosom J-P. Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *Journal of the Acoustical Society of America*. 2008; 124:2308–2319. [PubMed: 19062869]
9. Hillenbrand JM, Nearey TM. Identification of resynthesized /hvd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America*. 1999; 106:3509–3523. [PubMed: 10380673]
10. Hazan V, Markham D. Acoustic-phonetic correlates of talker intelligibility for adults and children. *Journal of the American Academy of Audiology*. 2004; 17:3108–3118.
11. Furui S. On the role of spectral transition for speech perception. *Journal of Acoustical Society of America*. 1986; 80:1016–1025.
12. Moon SJ, Lindblom B. Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America*. 1994; 96:40–55.
13. Bradlow AR, Krause N, Hayes E. Speaking clearly for children with learning disabilities: sentence perception in noise. *Journal of Speech, Language, and Hearing Research*. 2003; 46:80–97.
14. Hosom JP. Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*. 2009; 51:352–368. [PubMed: 20161342]
15. Sjölander K, Beskow J. WaveSurfer — an open source speech tool. *Proc. of ICSLP*. 2000:464–467.
16. Kain A, Hosom J-P, Niu X, van Santen J, Fried-Oken M, Staehely J. Improving the Intelligibility of Dysarthric Speech. *Speech Communication*. 2007; 49:743–759.
17. IEC/CD. Electroacoustics-sound level meters. 1996
18. Perkell JS, Zandipour M, Matthies ML, Lane H. Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues. *Journal of the Acoustical Society of America*. 2002; 112:1627–1641. [PubMed: 12398468]
19. Ferguson SH. Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *Journal of the Acoustical Society of America*. 2004; 116:2365–2373. [PubMed: 15532667]
20. Krause JC, Braida LD. Investigation alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *Journal of the Acoustical Society of America*. 2002; 112:2165–2172. [PubMed: 12430828]
21. Liu S, Rio ED, Bradlow AR, Zeng FG. Clear speech perception in acoustic and electric hearing. *Journal of the Acoustical Society of America*. 2004; 116:2374–2383. [PubMed: 15532668]
22. Helfer KS. Auditory and auditory-visual recognition of clear and conversational speech by older adults. *Journal of the American Academy of Audiology*. 1998; 9:234–242. [PubMed: 9644622]
23. Wingfield A, Poon LW, Lombardi L, Lowe D. Speed of processing in normal aging: Effects of speech rate, linguistic structure and processing time. *Journal of Gerontology*. 1985; 40:579–585. [PubMed: 4031406]

Figure 1. Weighting functions applied to the conversational formant contour (F1 through F4). Only HYB-MT and HYB-M require weighting functions.

Figure 2.

Formant contours (F1 through F4) of the word, "wheel" of three hybrid conditions. Dotted lines are conversational formant contours, and solid lines are the modified hybrid contours. Vertical dashed lines in *HYB-MT* and *HYB-M* represent phoneme boundaries. The circles and triangles on the formant contours were the average values obtained from clear speech.

Figure 3. Correct identification rates (%) for five conditions for each vowel in the perceptual experiment. Bars indicate standard errors for each condition. The significant differences between two conditions were shown with asterisks.

Figure 4.

Formant contour shape of $_{\text{CNV}}$ (red solid lines) and $_{\text{CLR}}$ speech (blue dotted lines).

Table 1

F0 values (Hz) and word duration (sec) used for the F0 contour template.

	/i/	/ɪ/	/e/	/ei/
1. onset of /w/	101.97	101.04	101.25	99.88
2. onset of /V/	124.91	123.43	117.91	119.11
3. Max F0	131.63	127.06	135.17	122.10
4. onset of /l/	95.51	102.13	98.49	90.76
5. offset of /l/	86.50	87.14	102.97	85.41
Mean F0	107.72	105.69	104.14	103.10
Word duration	0.42	0.38	0.40	0.45

Table 2

Formant steady-state (SS) values (Hz), F2 slope (Hz/ms) at vowel onset and offset of four vowels, and vowel duration (ms) in two speaking conditions. The difference (*D*) for each feature was computed by $_{CLR} - _{CNV}$.

	Styles	/i/	/ei/	/ɪ/	/e/
F1 SS values	CNV	375.59	498.34	457.96	581.24
	CLR	319.29	414.77	439.29	685.48
	<i>D</i>	-56.3	-83.57	-18.67	104.24
F2 SS values	CNV	1830.81	1601.06	1304.65	1215.76
	CLR	2439.31	2113.96	1724.17	1547.69
	<i>D</i>	608.5	512.9	419.52	331.93
F2 slope at vowel onset	CNV	16.5	11.96	8.13	6.74
	CLR	32.35	20.59	18.56	12.87
	<i>D</i>	15.85	8.63	10.43	6.13
F2 slope at vowel offset	CNV	-10.56	-10.06	-5.89	-3.94
	CLR	-15.83	-12.62	-6.79	-4.93
	<i>D</i>	-5.27	-2.56	-0.9	-0.99
Vowel duration	CNV	98.75	118.75	78.13	88.75
	CLR	229.38	236.25	141.88	151.88
	<i>D</i>	130.63	117.5	63.75	63.13

Vowels /i/ and /ei/ were found to be perceptually different between CNV and CLR speech (shown with asterisks).