THE EFFECT OF FORMANT TRAJECTORIES AND PHONEME DURATIONS ON VOWEL INTELLIGIBILITY

Akiko Amano-Kusumoto and John-Paul Hosom

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

ABSTRACT

We examined how much listeners can benefit from listening to "clear" (CLR) speech compared to "conversational" (CNV) speech, both spoken at different speaking rates. Vowel intelligibilities of four front vowels (/i:/, /I/, /E/, and /ei/) in background noise were measured with four speaking styles (CNV/SLOW, CNV, CLR, and CLR/FAST). Results showed only tense vowels of CLR speech had a significant difference between CNV and CLR speaking styles, after energy and F0 contour were normalized. We synthesized hybrid (HYB) speech whose formant features were equal to those of CLR speech, while all other features were taken from CNV speech. Primary conclusions from this study are (1) naturally-spoken fast CLR speech was not as intelligible as CLR speech, (2) enhancing formant frequencies to resemble those of CLR speech was effective at improving vowel intelligibility, and (3) spectral tilt and formant bandwidths were not contributing factors to the CLR speech benefit.

Index Terms— speech intelligibility, speech analysis, speech processing, speech enhancement

1. INTRODUCTION

The intelligibility of clear (CLR) speech, which is spoken deliberately clearly as if talking to a hard-of-hearing listener, is known to be higher than that of conversational (CNV) speech, which is spoken as if talking with a colleague [1]. A number of acoustic features have been recognized to be different between CNV and CLR speech [2, 3]. In this work, we focus on phoneme duration, formant steady-state (SS) values, and the formant transition regions. According to previous research, phoneme durations of CLR speech are longer, especially for tense vowels, and the vowel space (a two-dimensional representation of F1 and F2) of CLR speech is larger for the lax vowels, compared with CNV speech. Another study showed that formant SS values of vowels systematically undershoot for both CNV and CLR speech in the direction of formant frequencies of neighboring consonants, but the degree of undershoot is less for CLR speech in the context of /w/ and /l/ [3]. An increased rate of formant frequency change (or slope) is also observed in CLR speech.

It is known that the degree of formant undershoot depends on speaking style, word stress, vowel duration, and neighboring consonants [4]. However, it is not clear that observed formant undershoot in CNV speech is detrimental to vowel intelligibility. As shown in study [5], the formant transition region, where the slope is the greatest, is the most important region for syllable (consonant-vowel-consonant) identification.

This work was supported in part by NSF grant 0826654.

In this study, we first extend the study of Moon and Lindblom [3] by testing CNV and CLR speech spoken at a different speaking rates. We analyze the contribution of speaking style, speaking rate, and vowel identity to intelligibility levels.

Next, we determine whether acoustic features of formant SS values and/or transitions contribute to the improved intelligibility of CLR speech by creating hybrid (HYB) speech. HYB speech is, in this case, a synthetic speech signal that contains acoustic features from both CNV and CLR speech [6]. We modify formant trajectories of CNV speech to match the SS and transitions of CLR speech. In this preliminary work, we restrict our study to a single speaker. This research has the potential to be developed for assistive listening devices or diagnostic tools.

2. SPEECH CORPUS

As an extension of Moon's study [3], the four front vowels (/i:/, /I/, /E/ and /ei/) surrounded by the consonants /w/-/l/ were recorded in this study. The /wVl/ context with front vowels provides large second formant (F2) movement between consonants and the vowel due to coarticulation.

2.1. Speech Material

Four test words (*wheel, will, well*, and *wail*) in a carrier sentence were repeated 16 times each. The total of 64 sentences was randomized and the order of sentences was kept the same for each speaking style. The carrier sentence "it's easy to tell the size of a *WORD*" was used to facilitate the use of prosodic manipulation upon the elicitation of CNV and CLR speech at different speaking rates. The word of interest was equally stressed. Speech materials were spoken in four speaking styles (CNV/SLOW, CNV, CLR, and CLR/FAST).

2.2. Recordings

One male speaker, a native speaker of North-American English, recorded the speech materials in four recording sessions. The recording of CNV speech was followed by CLR speech in the first and second recording sessions, using the speaker's own distinction between CNV and CLR speech production. For the third session, CNV speech was spoken at a deliberately slow rate of the speaker's choice. For the fourth session, CLR speech was recorded at a fast speaking rate. A speaking rate other than natural is indicated after the speaking style, e. g. CLR/FAST and CNV/SLOW. It was not the goal of this study to match the speaking rate of CLR/FAST speech with CNV speech, or to match the rate of CNV/SLOW speech with CLR speech. The purpose was to have variety of speaking rates with CNV and CLR speaking styles. The averages of the resulting

Experiment 1.					
Speaking Styles	**/i:/**	**/I/**	**/E/**	**/ei/**	**Mean**
CNV/SLOW	91.1	64.4	81.1	70.0	76.7
CNV	15.6	60.0	75.6	13.3	41.1
CLR	95.6	81.1	85.6	100.0	90.6
CLR/FAST	81.1	81.1	88.9	71.1	80.6
Mean	70.8	71.7	82.8	63.6	
Experiment 2.					
Conditions	**/i:/**	**/I/**	**/E/**	**/ei/**	**Mean**
HYB-SS	64.6	51.0	64.6	49.0	57.3
HYB-FS	68.8	54.2	69.8	19.8	53.1
HYB-TRD	93.8	47.9	67.7	79.2	72.1
CNV	36.5	52.1	68.8	28.1	46.4
CLR	79.2	51.0	62.5	81.3	68.5
Mean	68.5	51.3	66.7	51.5	

Table 1: Percent correct rates for Experiment 1 and Experiment 2.

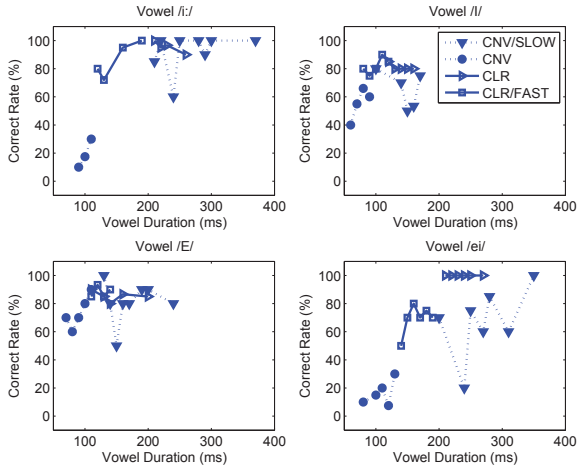


Fig. 1: Percent correct rates based on vowel duration for each vowel in Experiment 1.

speaking rates were 149 wpm, 365 wpm, 179 wpm, and 289 wpm for CNV/SLOW, CNV, CLR, and CLR/FAST, respectively.

3. EXPERIMENT 1: NATURALLY SPOKEN CNV AND CLR SPEECH

A perceptual experiment was conducted to examine whether the effects of CLR speaking style vary based on speaking rate and vowel identity. In order to examine the effect of phoneme duration and formant frequencies, the fundamental frequency (F0) contour and energy of the vowel was normalized across all speech samples. Isolated words, without a carrier sentence, were presented to the listeners.

Ten adults, aged between 19 and 38 years, were recruited for Experiment 1. All listeners were native speakers of North-American English with self-reported normal hearing.

3.1. Procedures

The perceptual experiment took place individually for each listener in a perceptual testing booth (*Whisper Room,* SE2000 series). A listener was seated in front of a computer monitor, listening to stimuli through circumaural headphones (Sennheiser HD 280 Pro), binaurally. A forced-choice test was used with four buttons, corresponding to the four choices ("**wheel**", "**will**", "**well**", "**whale**"), appearing on the user-interface screen.

12-talker-babble noise was used to simulate a noisy environment. The energy of the noise was adjusted to meet the desired SNR50-CNV level for each listener. The SNR50-CNV level refers to the signal-to-noise (SNR) ratio at which a listener can correctly identify the stimuli in the CNV speaking style 50% of the time. The listeners were tested in three sessions: the first two sessions were used for obtaining the listener's SNR50-CNV level, and the third session was for the vowel identification experiment. SNR50-CNV values from the second session were used for the vowel identification experiments, while the values from the first session were disregarded. The total of 144 stimuli (4 /wVl/ words $\times$ 4 speaking styles $\times$ 9 repetitions) were tested using a Latin Square design.

3.2. Results and Discussions

The results of Experiment 1 are shown in Table 1. Table 1 displays percent correct rates for each vowel identity and speaking style, averaged over 10 listeners. The average noise level (SNR50-CNV) was -1.08 dB (std: 2.11). Percent correct rates were converted to rationalized arcsine units (RAUs) prior to statistical analysis [7]. The effect of vowels (/i:/, /I/, /E/ and /ei/) and speaking styles (CNV/SLOW, CNV, CLR and CLR/FAST) were submitted to a two-way repeated-measures analysis of variance (ANOVA).

The results of two-way ANOVA (vowels $\times$ speaking styles) show that the main effects of vowels and speaking styles were significant ($p = 0.001; p < 0.0001$). For the tense vowels (/i:/ and /ei/), CLR speech was significantly more intelligible than CNV speech (both $p < 0.01$). On the other hand, for lax vowels (/I/ and /E/) the effects of speaking style were not significant (both $p > 0.01$). The confusion patterns in CNV speech, which had the least intelligibility compared with any other speaking style, showed that listeners tended to perceive the word "*wheel*" as "*will*" (73.33%), and the word "*whale* as "*well*" (44.44%) and "*will* (40.00%). In general, with babble background noise at the SNR50-CNV level, tense vowels with short vowel durations were more often perceived as lax vowels, while long tense vowels tended to be identified correctly.

The speaking rate affected intelligibility, showing that CLR/FAST speech is less intelligible than CLR speech, and that CNV speech is less intelligible than CNV/SLOW speech. This indicates that the faster speaking rates resulted in less intelligible speech. It may not be possible to obtain a CLR speech benefit at fast speaking rate for naturally produced speech, as shown in [8]. Figure 1 shows percent correct rates based on the vowel duration of the stimulus. It is clearly shown that the shorter vowel durations have less percent correct rates. The one exception is in the CNV/SLOW data, which shows a notch at one duration per vowel. This seems to be an anomoly that is not indicative of the underlying trend. For CNV speech, it is uncertain whether the cause of less intelligible tense vowels is due to short vowel duration or the large amount of formant undershoot. In Experiment 2, we will examine whether it is possible to improve vowel intelligibility by modifying formant frequencies.

The formant steady-state (SS) values, extracted at the midpoint of each vowel, were weakly correlated with vowel intelligibility ($\rho = 0.2528$, $p = 0.0013$). A not strong but significant correlation was found for both vowel duration and word duration with vowel intelligibility ($\rho = 0.4208$, $p < 0.0001$; $\rho = 0.4748$, $p < 0.0001$). The results of correlation analysis do not imply that higher F2 frequency or longer duration are the cause of improved vowel intel-

ligibility. In Experiment 2, we will examine the cause of improved intelligibility of CLR speech by creating HYB speech.

4. HYBRIDIZATION ALGORITHM

The hybridization (HYB) algorithm proposed here is a signal processing technique that modifies certain acoustic features of CNV speech to match those of CLR speech [6]. The results from acoustic analysis revealed that CNV and CLR speech had inherently different F2 SS values and different F2 slopes between phonemes /w/ and /V/. Perceptual experiment 1 showed that the vowel intelligibility for /i:/ and /ei/ was higher for CLR speech than CNV speech and that short durations negatively impacted the intelligibility of tense vowels. These results motivated us to examine whether intelligibility could be improved by reducing degree of formant undershoot, or whether CNV speech with modified formants is inherently less intelligible than CLR speech because of the short duration of CNV speech. Our hypothesis was that formant SS values (and possibly transitions) of CLR speech contribute to improved intelligibility.

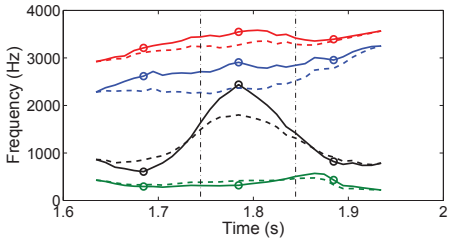
The first step of the hybridization algorithm was to extract *target* formant SS and/or formant transition values from CLR speech. Then, the HYB formant *trajectories* with *target* values were designed. The third step was to modify formant values of CNV speech by analysis and synthesis methods to match the target formant trajectories.

4.1. HYB Conditions

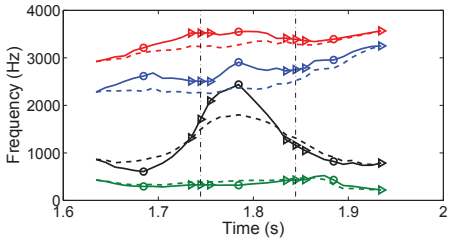
Three HYB conditions to test the effect of SS values, transitions, or vowel duration (HYB-SS, HYB-FS, and HYB-TRD) were evaluated. The term HYB-SS indicates that formant SS values of HYB speech are those of CLR speech. Similarly, the term HYB-FS indicates that SS values and formant transitions at phoneme boundaries of HYB speech are those of CLR speech. Finally, the third condition HYB-TRD indicates that the entire formant trajectory (not only SS values or transitions) and phoneme durations of HYB speech are those of CLR speech. The reason to include phoneme durations in HYB-TRD is because our previous study showed that changing the combination of short-term spectra and phoneme durations improved the sentence intelligibility over that of the CNV speech [6]. Also, HYB-TRD provides a test of the quality of our hybridization method with formant modification.

4.1.1. HYB-SS: CLR steady-state values

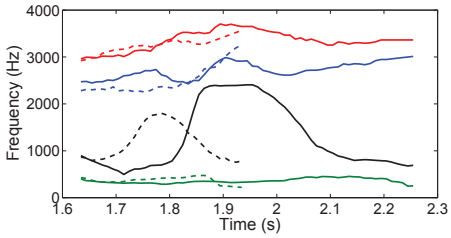
The *target* SS values were extracted at the midpoints of each phoneme /w/, /V/ and /l/ of CLR speech, and averaged over 16 samples per word. The process of designing a HYB-SS formant trajectory required a weighting function for each formant trajectory (F1 through F4). The weighting function was designed to be linearly ascending or descending to describe the ratio between the *target* SS values and the CNV SS values at the midpoint of each phoneme, using phoneme duration from CNV speech. 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 signal. Then, the original CNV formant trajectory was shifted by the weighting function to obtain the HYB-SS formant trajectories. The resulting HYB-SS formant trajectory has *target* SS values from CLR speech. Figure 2(a) shows the original CNV (dashed line) and HYB-SS formant trajectory (solid line) after the weighting function was applied.



(a) Formant Trajectories of HYB-SS



(b) Formant Trajectories of HYB-FS



(c) Formant Trajectories of HYB-TRD

Fig. 2: Formant trajectories (F1 through F4) of the word, "*wheel*". Dotted lines are CNV formant trajectories, and solid lines are the modified trajectories. Vertical dashed lines in (a) and (b) represent phoneme boundaries. The circles and triangles on the formant trajectories indicate the SS values and the transitions, respectively.

4.1.2. HYB-FS: CLR steady-state values and formant transitions

In addition to the *target* SS values, *target* transitions at phoneme boundaries over a 20 ms range were extracted and averaged over 16 samples of CLR speech per word. In designing the weighting function to include phoneme transitions, the ratio was calculated between *target* values and CNV formant values (F1 through F4) at midpoints of the phonemes, and at three points near the phoneme boundaries for both /w/ to /V/ and /V/ to /l/. Similar to the previous condition HYB-SS, the weighting function for each formant was designed to be linearly ascending or descending to describe the desired ratio. Then, the original CNV formant trajectory was shifted by the weighting function. In this way, the formant trajectories of HYB-FS were guaranteed to have SS values and transitions at the phoneme boundaries that were identical with *target* values. The formant trajectories of HYB-FS (solid line) and the original CNV (dashed line) are shown in Figure 2(b).

4.1.3. HYB-TRD: CLR formant trajectories with phoneme durations

Unlike HYB-SS and HYB-FS conditions, the process of designing HYB-TRD formant trajectories did not require a weighting function. Because phoneme durations were modified to match CLR speech at the synthesis stage, formant trajectories from CLR speech

were copied as HYB-TRD formant trajectories. As shown in Figure 2(c), the formant trajectories of HYB-TRD (solid line) have longer phoneme durations than the original CNV (dashed line).

4.2. Speech synthesis with HYB formant trajectories

For all HYB conditions, after new formant trajectories were designed, the original CNV speech was analyzed, the existing formants were removed by inverse filtering, and the HYB speech was synthesized with new formant trajectories.

First, the speech waveform was analyzed with a pitch-synchronous frame that spans two pitch periods with a one pitch period overlap. Four resonant frequencies (F1 through F4) in one frame were removed by applying an inverse filter that was designed with formant frequency values from CNV speech. The residual signal from inverse filtering contained primarily the glottal source and higher formants.

The new formant trajectories (F1 through F4) were used to design all-pole digital filters acting as vocal tract filters. The bandwidths of each filter were unchanged from those of the original CNV speech. The speech waveform in each frame was obtained by applying the all-pole digital filters to the residual signal.

For HYB-TRD, it was required to stretch the phoneme duration to match that of CLR speech. At the synthesis stage, residual signals were repeated as necessary to obtain the desired phoneme durations. The HYB-SS and HYB-FS conditions had no duration modification. In all cases, overlap-add was used to create the final waveform.

5. EXPERIMENT 2: SYNTHETIC HYB SPEECH

A perceptual experiment was conducted to examine whether HYB speech with CLR speech formant values, with and without stretching phoneme durations, is more intelligible than CNV speech. Six adults, aged between 19 and 34 years, participated in Experiment 2.

5.1. Procedures

Four vowels (/i:/, /I/, /E/ and /ei/) with /wVl/ contexts were tested in 12-talker-babble noise at each listener's SNR50-CNV level. The stimuli used in Experiment 2 were the three types of HYB speech (HYB-SS, HYB-FS, and HYB-TRD) and the original CNV and CLR speech as a baseline. All stimuli had a normalized energy value and F0 contour. The experimental procedure was the same as in Experiment 1 (Section 3.1). A total of 320 stimuli (4 /wVl/ words $\times$ 5 stimuli type $\times$ 16 repetitions) were presented to each listener.

5.2. Results and Discussions

The results of Experiment 2 are shown in Table 1. The average SNR50-CNV used for the background noise was -4.44 dB (std: 0.99). Percent correct rates were converted to the rationalized arcsine units (RAUs) prior to statistical analysis [7]. The planned t-test revealed that the difference in vowel intelligibility between CNV and CLR speech was significant for tense vowels (both $p < 0.001$). For the vowel /i:/, the intelligibility differences between CNV and all three HYB conditions were significant (HYB-SS, $p = 0.0043$; HYB-FS, $p = 0.0016$; HYB-TRD, $p < 0.0001$). For the vowel /ei/, only HYB-SS and HYB-TRD speech improved vowel intelligibility over CNV speech ($p = 0.0388$, $p < 0.001$). The difference in results between HYB-SS and HYB-FS for the vowel /ei/ suggests the importance of the phoneme transition region; however, it is not yet clear how the transitions should be modified to maximize intelligibility. Because of the lack of intelligibility difference between CNV and CLR for lax vowels, there was no room for the formant modification to improve the intelligibility of CNV speech.

The confusion patterns showed that "*wheel*" in CNV speech was perceived as "*will*" 39.58% of the time, while the correct response was 36.46%. The "*wheel*"-"*will*" confusion was improved to 26.04%, 22.92%, and 4.17% for HYB-SS, HYB-FS, and HYB-TRD, respectively. The word "*whale*" was often confused with "*will*", at 40.63%, 25.00%, and 32.29% in CNV, HYB-SS and HYB-FS conditions. However, "*whale*" was perceived as "*wheel*" in HYB-FS (30.21%) much more than in HYB-SS (12.50%). Similar to the results in Experiment 1, the short tense vowels were more often perceived as lax vowels. It is worth noting that the vowel /ei/ was confused with /i:/ in HYB conditions, even at short durations.

In summary, for the vowels /i:/ and /ei/, even with short durations, the formant modification targeting CLR SS values was effective to significantly improve vowel intelligibility. The HYB-TRD results confirm that the hybridization algorithm can yield high-quality speech when modifying formant frequencies. It can also be concluded that spectral tilt and formant bandwidth were not important contributions to the improved intelligibility of CLR speech for our normal hearing listeners and this speaker.

6. CONCLUSION

We draw four conclusions from this study: (1) It may not be possible to obtain a CLR speech benefit in naturally-spoken fast speech. (2) The HYB algorithm can successfully increase the intelligibility of CNV speech to CLR intelligibility levels by formant and duration modification (HYB-TRD). (3) Vowels with short durations can have significantly improved intelligibility by formant modification. (4) The spectral tilt and formant bandwidth did not contribute to improving intelligibility. In the future, we plan to study HYB speech with formant targets from isolated vowels, to further evaluate the impact of SS and transitions.

7. REFERENCES

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

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

[3] S. J. Moon and B. Lindblom, "Interaction between duration, context, and speaking style in English stressed vowels," *Journal of the Acoustical Society of America*, vol. 96, no. 1, pp. 40–55, 1994.

[4] B. Lindblom, "Spectrographic study of vowel reduction," *Journal of the Acoustical Society of America*, vol. 35, no. 11, pp. 1773–1781, 1963.

[5] S. Furui, "On the role of spectral transition for speech perception," *Journal of Acoustical Society of America*, vol. 80, no. 4, pp. 1016–1025, 1986.

[6] A. Kain, A. Amano-Kusumoto, and J.-P. Hosom, "Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility," *Journal of the Acoustical Society of America, accepted for publication, Oct.*, 2008.

[7] G. A. Studebaker, "A 'rationalized' arcsine transform," *Journal of Speech and Hearing Research*, vol. 28, pp. 455–462, 1985.

[8] J. C. Krause and L. D. Braida, "Investigation alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility," *Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2165–2172, 2002.