Can modified casual speech reach the intelligibility of clear speech?

M. Koutsogiannaki[1], *M. Pettinato*[2], *C. Mayo*[3], *V. Kandia*[1] *and Y. Stylianou*[1]

[1]Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
[2] Speech, Hearing and Phonetic Sciences, University College London, UK
[3] Centre for Speech Technology Research, the University of Edinburgh, UK
{mkoutsog, vkandia, yannis}@ics.forth.gr, m.pettinato@ucl.ac.uk, catherin@ling.ed.ac.uk

Abstract

Clear speech is a speaking style adopted by speakers in an attempt to maximize the clarity of their speech and is proven to be more intelligible than casual speech. This work focuses on modifying casual speech to sound as intelligible as clear speech. First, we examine the role of speaking rate for intelligibility. Clear and casual speech signals are time-scale stretched, matching the average duration of the casual and clear speech respectively. Next, spectral shaping and dynamic range compression are considered for increasing the loudness of the original casual speech while keeping the power of signals unaffected. Subjective tests with speech-in-noise conditions using speech shaped noise at -3, 0, and 5 dB SNR show that clear speech with high speaking rate is less intelligible than the original clear speech but still more intelligible than the unmodified casual speech. However, the intelligibility score for the time-scaled modified casual speech is deteriorated. In contrast, the loudness amplification considerably improved the intelligibility of the casual speech, reaching the scores of original clear speech. Objective measurements based on Speech Intelligibility Index (SII) are well correlated with the subjective test except for the time-scaled casual signal.

Index Terms: Clear speech, Casual speech, Speech intelligibility, Spectral modifications

1. Introduction

In difficult communication situations, talkers modify their speech: they typically speak more slowly, more loudly and carefully articulate sounds in order to successfully deliver their message. The speaking style that they adopt depends on the communication difficulty; speaking against a background of noise (Lombard speech), to a hearing impaired listener or to a non-native language speaker. A common characteristic between these different styles of speech is their greater intelligibility than undisturbed natural speech, namely casual speech. In this work, clear speech is examined. Clear speech is a speaking style adopted by speakers when they are instructed to maximize the clarity of their speech. The goal is to modify casual speech to sound as intelligible as clear speech.

Analysis on casual and clear speech signals has shown differences in acoustic and phonetic level of speech. Clear and casual speech may differ among others in intensity [1], [2], speaking rate [1], [2], number and duration of pauses [1], [2], pitch [3], [4], long term RMS spectra [1], [5], [4], modulation spectra [5], vowel duration and vowel space [1], [6], [3], [4]. However, analysis on databases of clear and casual signals on various studies showed that these observed differences are not present to all speakers. For example, speakers can elicit clear speech with and without changing their pitch or increasing their voice volume level. A recent study suggests that clear speech can also be produced without decreasing the speech rate, after training the speakers [5]. This suggests that clear speech has inherent acoustic properties independent of rate that contribute to improved intelligibility.

In this work, clear speech is transformed to match casual speech and casual speech to match clear speech in terms of duration. This is performed in order to check the intelligibility advantage of the low speaking rate in clear speech perception. Indeed, time-scaled clear speech in higher speaking rates has lower intelligibility according to subjective and objective evaluations. This suggests that speaking rate plays a significant role in intelligibility. However, casual speech scaled on lower speaking rates is observed to have decreased intelligibility, instead of providing higher intelligibility scores. Therefore, time-scaling seems an inappropriate method for enhancing the intelligibility of casual speech. To that purpose, Spectral Shaping (SS) and Dynamic Range Compression (DRC) are implemented to enhance intelligibility of casual speech in noisy and noise-free conditions [7]. Subjective and objective measure tests are performed on five sets of signals; namely, on the initial database of clear and casual signals and, additionally, on the time-scaled casual signals on lower speaking rates, the time-scaled clear signals on higher speaking rates, and the SS-DRC modified casual signals.

This paper is organized as follows. Section 2 describes the database of clear and casual speech signals [4]. Section 3 explains the methodology of adjusting the duration of clear speech to match the duration of casual speech and vice versa, and of transforming the casual speech with the SS-DRC method. Section 4 briefly describes the setup of the listening tests and presents the results based on subjective and objective evaluations. Finally, Section 5 concludes the paper.

2. The Database

The clear and casual speech used for analysis is the read speech from the LUCID database. Read speech is an exaggerated form of clear speech relative to the spontaneous clear speech [4]. The LUCID dataset consists of read clear and read casual signals uttered by speakers who performed two tasks. In the first task they were asked to read the sentences “casually as if talking to a friend” whereas in the second task they were instructed to speak “clearly as if talking to someone who is hearing impaired”. Speakers in this database are Southern British English between 19 and 29 years old with no speech or language disorders. The sentences are meaningful and simple in syntax. From the corpus of the LUCID database, 70 distinct sentences are selected, uttered by 14 female speakers and 9 male speakers.

Analysis on the database [4] showed that clear elicited speech differs from casual speech in pitch, word duration, energy level between 1 and 3 kHz in the long-term spectrum, and in vowel formant range. Specifically, speakers produced speech with higher fundamental frequency, higher frequency energy in band $1-3$kHz and higher range in first and second formant in clear speech than casual speech. Moreover, speakers slowed down their speech to a greater extent when reading clearly.

From pilot listening tests and objective measurements, pitch modifications seem not to contribute to speech intelligibility. Previous studies report that lower speaking rate greatly contributes to the clarity of clear speech. To examine whether this applies or not, we exclude the duration factor by modifying the clear signals to match the duration of the casual signals. If duration indeed plays a role in intelligibility, then the most significant modification of the casual speech should be time scale modification to lower speaking rates.

3. Methodology

A simple method is implemented in order to find which features contribute to the intelligibility advantage of clear speech. First, the 69 distinct clear sentences (set A) and the same utterances uttered by the same speakers in a casual style (set B) are scaled in time. This approach is implemented to test the contribution of speaking rate to the intelligibility of clear speech. Clear sentences are time-scaled to match the duration of casual sentences (set C), and casual sentences are time- scaled to match the duration of clear sentences (set D). For set E, spectral transformations are performed on casual signals. These consist of Spectral Shaping and Dynamic Range Compression. The former reduces the bandwidth of the formants by increasing the distance between peaks and valleys, whereas the latter reallocates the energy of the signal, by shifting it from high-energy into lower-energy parts.

The five sets are evaluated both subjectively and objectively in the presence of noise. A subjective evaluation includes listening tests which are performed by native and non-native listeners, while an extended version of the Speech Intelligibility Index (ESII) is introduced for objective evaluation.

3.1. Time Alignment and Time-Scale Modification

A preprocessing is performed on the dataset to remove low-pass noise from breath and lip effects, using a 5-order low pass digital elliptic filter with 80Hz cut-off frequency. Then, time alignment is performed by hand at the segmental level. The time-alignment information is used by the Waveform Similarity Based Overlap-Add algorithm (WSOLA, [8]) that time scales the one signal to match the duration of the other signal.

3.2. Spectral Shaping and Dynamic Range Compressing

The goal of Spectral Shaping is to increase the "crisp" and "clean" quality of the speech signal, and therefore improve the intelligibility of speech even in clear (noise-free) conditions [7]. For this, both adaptive and fixed spectral shaping operators are used. The adaptive spectral shaping takes into account the probability of voicing given a speech frame, while the fixed spectral shaping is independent of the probability of voicing. The adaptive spectral shaping consists of (i) adaptive sharpening where the formant information is enhanced, and (ii) an adaptive pre-emphasis filter. The adaptive (to the probability of voicing) characteristic of the suggested system is important for not introducing artifacts, in the processed signal especially in fricatives, silence or other "quiet" areas of speech. The purpose of the fixed (non-adaptive) spectral shaping is to protect the speech signal from low-pass operations during the reproduction of the signal.

The output of the Spectral Shaping system is the input to the Dynamic Range Compressor (DRC). DRC has a dynamic and a static stage. During the dynamic stage, the envelope of the signal is dynamically compressed with 2ms release time constant and almost instantaneous attack time constant. The signal envelope is based on the Hilbert transform and a moving average operator with order determined by the average pitch of speakers gender. After the dynamic compression of the signal envelope, a static amplitude compression is applied. During the static amplitude compression, the 0 dB reference level is a key element in forming the Input/Output Envelope Characteristics (IOEC). For the current system this was set to 0.3 of the peak of the signal. The whole system is based on a frame-by-frame analysis and synthesis. In each frame the magnitude spectrum is computed using FFT and then manipulated in the way mentioned above. Overlap and add is then used to reconstruct the modified signal. The whole process is very fast and can run in real time.

4. Results

In this section, both subjective and objective evaluations are presented.

4.1. Subjective evaluations

In the perceptual tests, speech shaped noise is added to the signals to create the test signals, with Signal to Noise Ratio (SNR) of $\{-3,\ 0,\ 5\}$ dB. Therefore, for the five set of signals $\{$A,B,C,D,E$\}$ and for the 3 different SNRs, a dataset of 5x3 test signals is created. From this dataset, each listener randomly hears signals with the limitation of hearing each sentence only once. Then, the listener evaluates a sentence based on the description in Table 4.1.

The listening test was performed by 24 native listeners and 15 non-native speakers and the corresponding subjective scores are depicted in Fig. 1(a) and Fig. 1(b). The scores from 1-5 are scaled from 0-1.

Score	Description
5	if you understood the whole sentence
4	if you understood the sentence except one or two words
3	if you could barely understand the sentence
2	if you could understand some words but not the message
1	if you could not understand anything at all

Table 1: Subjective Scores Description

The acoustic analysis results show that in low noisy environment ($SNR = 5$dB), clear speech is still more intelligible than casual speech both for native (7%) and for non-native listeners (17%), where the intelligibility advantage of clear speech is much greater in non-native listeners. When time scaling the clear speech to match the duration of the casual speech, the intelligibility is reduced 5% for native listeners and 8% for non-native listeners in low SNR condition ($SNR = 5$dB). However, in high presence of noise ($SNR = -3$dB) eliminating the low speaking rate property of the clear speech reduces its intelligibility to 18% for native listeners and 11% to non-native listeners. Therefore, it is verified that speaking rate plays a significant role in the intelligibility of the clear speech.

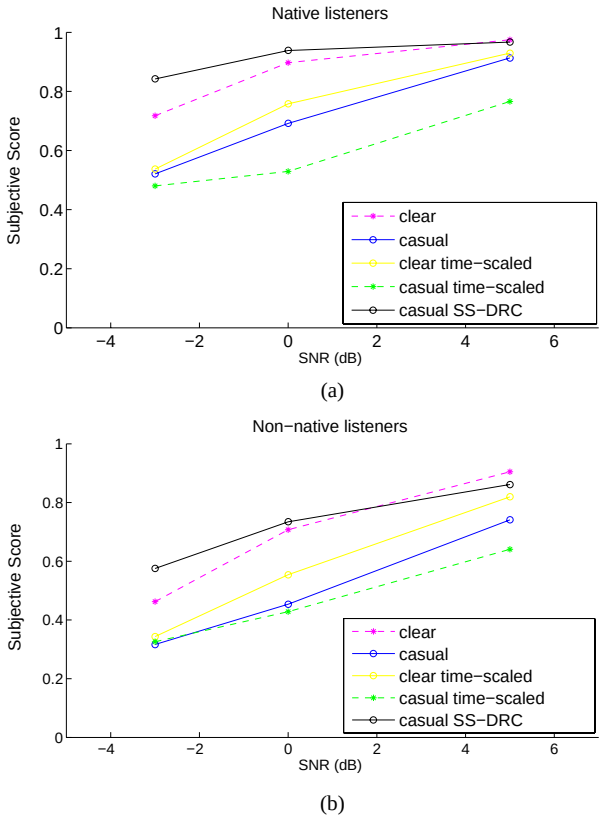


Figure 1: Subjective Measure Score for the 5 set of signals for different levels of SNR. a) Native listeners b) Non-native listeners

Time-scaled clear speech, however, seems to still be more intelligible than casual speech for non-native listeners (Fig. 1(b)). This suggests that clear speech has other acoustic properties, despite rate that contribute to its improved intelligibility. This assumption is reinforced by the fact that the time-scaled casual speech on lower speaking rates gives lower intelligibility scores than the unprocessed casual speech, with a more negative impact to the native listeners. Native listeners actually reported that the time-scaled casual signals are irritating. This can be justified by the fact that expanding a casual signal does not mean that its missing information is replaced. For example, Fig. 2 depicts the waveform of the phrase "full of". The clear signal, apart from the extended duration of the phonemes, contains also pauses. The casual signal, however neither contains pauses nor all the acoustic information that the clear signal has. The expansion of the signal in that case may reduce intelligibility.

Transforming the casual signals in the spectral domain significantly raises intelligibility not only in a high noise environment but also in low noise levels (Fig. 1(a), Fig. 1(b)). Native listeners reported a 32% increase of intelligibility after SS-DRC, whereas non-native listeners reported a 27% raise. For low SNR levels (-3dB) the transformed SS-DRC casual speech is 11% more intelligible than clear speech, as Fig. 1(a) and Fig. 1(b) depict, while for high SNR values (5dB), modified SS-DRC casual speech and clear speech share the same intelligibility score.

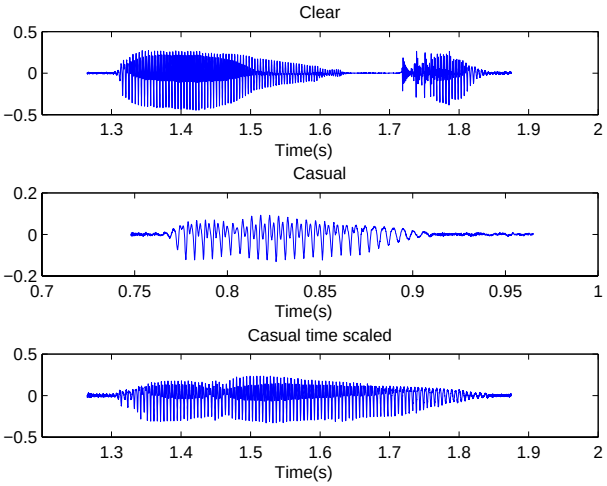
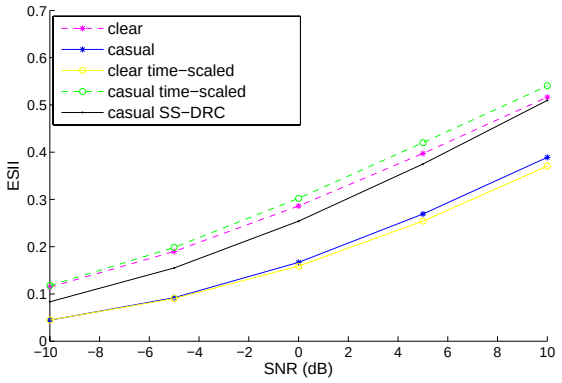


Figure 2: Casual signal time-scaled to match the duration of the clear signal

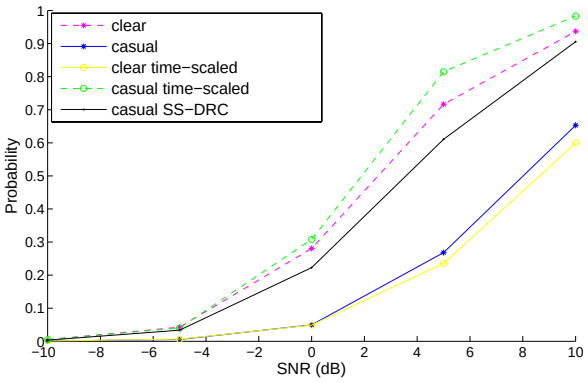
4.2. Objective evaluations

Objective measure tests were also performed, using a modified version of the extended Speech Intelligibility Index (ESII). For the computation of SII we followed the steps described in [9], towards what is referred to as Extended SII. First, an FIR Filter Bank is used to filter speech and noise signals into 21 critical bands [10]. Each filter in the filter bank is a linear FIR filter of type I and order 200. Next, the time varying intensity of the signal is computed for each output of the filter bank. For this, non overlapped rectangular windows are used with window lengths ranging from 35 ms at the lowest band (center frequency 50 Hz) to 10 ms at the highest band (center frequency 7000 Hz). The windows are aligned such that they end simultaneously [9]. The intensity level is normalized to dBSPL using the absolute threshold of hearing (10^{-12} watts). At a given instant, the instantaneous SII is computed following a standard procedure (ANSI $S3.5-1997$, [11]) using the so-called speech perception in noise (SPIN) weighting function and the estimated speech and noise normalized intensities. Finally, the SII for a speech-in-noise condition is determined by simple averaging across all the instantaneous SII values. The objective intelligibility score computed by ESII was successfully validated using results from a listening test described in [12].

According to the ESII measure, clear speech and casual SS-DRC speech have higher intelligibility scores than casual speech (Fig.3(a)) with higher probability (Fig.3(b)) of identifying a sentence for SNR levels above -5dB. On the other hand, casual speech and time-scaled clear speech that have the same duration give the same score of ESII independent of the SNR level (Fig.3(a)). This agrees with the subjective evaluations even if the scores that the ESII gives for the signals in the specific SNRs are much lower. Objective measures give contradictory results with the subjective scores in the case of the time-scaled casual speech. Obviously, the ESII accounts for spectral differences and not for the linguistic content of a sentence which may be missing in casual speech.



(a)



(b)

Figure 3: Objective Measure Score for the five set of signals for different levels of SNR. a) Extended Speech Intelligibility Index b) Probability of correctly identifying a sentence

5. Discussion

Detecting the features that make clear speech more intelligible than casual is not an easy task. The trend which appears in the majority of speakers is the lower speaking rate of clear speech compared to casual speech.

Eliminating the duration factor from the clear speech, verifies that clear speech is more intelligible than casual speech due to its duration. However, duration is not the only contributing factor. Subjective tests showed that the time-expanded casual speech had lower intelligibility scores than the unprocessed casual speech. The time-expansion cannot fill the gap of the missing phonetic-level and acoustic-level information; clear speech contain pauses and phonemes often eliminated on casual speech.

Therefore, the question is how can we enhance the intelligibility of the casual signals when the features that improve intelligibility are still under consideration, and those known to improve intelligibility cannot be directly applied to casual speech? This work answers this question by applying Spectral Shaping and Dynamic Range Compression to casual speech. Subjective evaluations show that modified casual speech gives greater intelligibility scores than clear speech in the presence of high level of noise and similar intelligibility scores in high SNR.

6. Acknowledgments

The authors would like to thank Vasilis Karaiskos and Valerie Hazan for organizing the listening tests in the University of Edinburgh and in the University College of London, respectively. This work is part of the the Listening Talker project (LISTA), funded under the EU Framework 7 Future and Emerging Technologies (FET) Programme.

7. References

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

[2] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing iii: an attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech.," *Journal of Speech and Hearing Research*, vol. 32, pp. 600–603, 1989.

[3] A. R. Bradlow, N. Kraus, and E. Hayes., "Speaking clearly for learning-impaired children: sentence perception in noise.," *Journal of Speech, Language, and Hearing Research*, vol. 46, pp. 80–97, 2003.

[4] V. Hazan and R. Baker, "Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style?," *DiSS-LPSS*, pp. 7–10, 2010.

[5] J. Krause and L. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *JASA*, vol. 115, no. 362-378, 2004.

[6] S. H. Ferguson and D. Kewley-Port., "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners.," *Journal of the Acoustical Society of America*, vol. 112, pp. 259–271, 2002.

[7] B. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Trans. Audio Acoust.,*, vol. 17, no. 1, 1969.

[8] M. Demol, K. Struyve, W. Verhelst, H. Paulussen, P. Desmet, and P. Verhoeve Author, "Efficient non-uniform time-scaling of speech with wsola for call applications," *Proceedings of InSTIL/ICALL2004 NLP and Speech Technologies in Advanced Language Learning Systems*, Venice 17-19 June, 2004.

[9] K. S. Rherbergen and N. J. Versfeld, "Speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *JASA*, pp. 2181–2192, 2005.

[10] Simon Fraser, *Handbook for Acoustic Ecology*, University and ARC Publications, 2 edition, 1999.

[11] ANSI S3.5-1997, "American national standard methods for calculation of the speech intelligibility index," Tech. Rep., American National Standards Institute, New York, ANSI (1997).

[12] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified hmm-based synthetic speech in noise?," *Interspeech 2011*, Florence, Italy, 2011.