Assessing the Intelligibility Impact of Vowel Space Expansion via Clear Speech-Inspired Frequency Warping

E. Godoy, M. Koutsogiannaki, Y. Stylianou

Foundation of Research and Technology Hellas, Institute of Computer Science, Crete, Greece

Abstract

Among the key acoustic features attributed with the intelligibility gain of Clear speech are the observed reduction in speaking rate and expansion of vowel space, representing greater articulation and vowel discrimination. Considering the slower speaking rate, previous works have attempted to assess the intelligibility impact of time-scaling casual speech to mimic Clear speech. In a complementary fashion, this work addresses the latter of the key traits observed in Clear speech, notably vowel space expansion. Specifically, a novel Clear speech-inspired frequency warping method is described and shown to successfully achieve vowel space expansion when applied to casual speech. The intelligibility impact resulting from this expansion is then evaluated objectively and subjectively through formal listening tests. Much like the relevant time-scaling works, the frequency warping that expands vowel space is not shown to yield intelligibility gains. The implications are thus that further analyses and studies are merited in order to isolate the pertinent acoustic-phonetic cues that lead to the improved intelligibility of Clear speech.

Index Terms: Speech Intelligibility, Clear speech, Vowel Space Expansion, Frequency Warping

1. Introduction

When faced with adverse communication conditions, human beings adopt speaking styles to increase intelligibility. Clear speech represents such a style that speakers use when a listener faces a communication barrier [1, 2, 3, 4]. Among the acoustic modifications associated with Clear speech, decreased speaking rate and vowel space expansion are often attributed with established gains in intelligibility [4, 3, 5, 6]. However, in the case of decreased speaking rate, attempts to time-scale speech to slow it down have not proven effective at increasing intelligibility [7, 8]. Alternatively, the present work focuses on the vowel space expansion observed in Clear speech. In a fashion similar to works addressing time-scaling modifications imitating those of Clear speech, the following work seeks to modify speech to expand vowel space and then assess the resulting intelligibility impact.

Though vowel space expansion is often largely attributed with the intelligibility advantage of Clear speech [1, 4, 6], the gains resulting specifically from this expansion unfortunately remain obscure, as reliably manipulating vowel spaces is a challenging feat discouraged by limitations in accurate formant estimation and treatment. Accordingly, speech modifications aiming to exploit the stipulated intelligibility gains of vowel space expansion are essentially non-existent. That said, in related work seeking to increase natural and synthetic speech intelligibility, a general upwards shift of formants proved beneficial in averting a low-frequency noise masker [9, 10], though this benefit was not present for other noise types. Additionally, the recent work in [11] used a common Gaussian Mixture Model (GMM) based voice conversion technique to statistically transform formant frequencies and the spectral envelope of casual vowels to resemble those of Clear. However, conducted intelligibility tests only considered the spectral envelope transformation, thus evaluating more than formant movement, including spectral tilt (e.g., amplitude) modifications. In sum, to the authors' knowledge, there has been no previous effort to address vowel space expansion in isolation and to asses the corresponding intelligibility impact. The present work focuses on this task, adopting an original approach that side-steps explicit formant detection and statistical learning, specifically expanding vowel space via frequency warping inspired by Clear speech analyses.

Typically used in voice conversion [12, 13, 14] or speech recognition (e.g., VTLN), frequency warping is employed here in a novel manner as a means for vowel space expansion. In this work, both the application and algorithm estimation for the frequency warping are original. Fundamentally, the appeal of frequency warping for vowel space expansion is that it offers a way of shifting speaker formants, while both avoiding notable speech degradations and limiting dependence on accurate formant detection. In particular, this work proposes a frame-based piecewise linear frequency warping function. Unlike the related Dynamic Frequency Warping (DFW) algorithms used for voice conversion [13, 14], however, the intervals of the warping function are defined in this work by sampling a curve (based on spectral peak locations) of exaggerated formant shifts that is drawn from Clear-casual vowel space analyses. In examining the vowel space of warped casual speech, it is confirmed that the proposed approach successfully yields expansion. Then, the corresponding intelligibility impact is assessed using an objective index and formal listening tests. In the end, results ultimately motivate more careful consideration and qualification of the Clear speech intelligibility advantage in relation to vowel space expansion.

This article is structured as follows. Section 2 illustrates the Clear speech vowel space expansion and corresponding formant shifts. Section 3 then describes the proposed frequency warping algorithm and demonstrates the resulting vowel space expansion. Section 4 evaluates the intelligibility impact of the proposed approach. Finally, Section 5 concludes.

2. Clear Speech Vowel Space Analyses

2.1. Speech Corpora and Vowel Space (V.S.) Generation

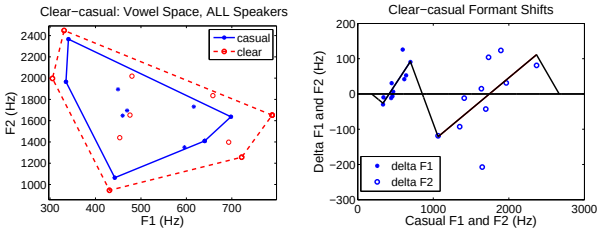
The speech data used for the Clear (and casual) vowel space analyses is read speech of the LUCID database [3]. In the LUCID database, speakers were asked to read meaningful and syntactically simple sentence in two ways: 1) "casually as if talking to a friend," 2) "clearly as if talking to someone who is hearing impaired." This work examines 8 (Southern) British

English speakers (4 male, 4 female) with 50 distinct sentences per speaker. The speech sampling rate is 44.1kHz and all of the speech was segmented using an HTK-based audio-to-text aligner (provided by University College London), without manual corrections.

The vowel spaces in this work are generated as follows. First, formant analysis is performed using Praat, which exploits the Burg algorithm [15]. The representative pair of F1 and F2 values for each vowel instance is then taken as the values at the center of the speech segment. For each vowel, the mean over all of the vowel instances is trimmed, with 95% of the data kept, in order to limit the influence of potential outliers. Then, the convex hull (i.e., a polygon fit that encompasses all of the data points) is used to represent the vowel space area, as it effectively captures the maximal area that the points in the vowel space span.

2.2. Observed V.S. Expansion and Formant Shifts

Fig.1a shows the vowel spaces calculated using all of the vowel instances for the speakers in the specified LUCID corpora. It is evident from Fig.1a that the Clear speech vowel space is expanded compared to the casual speech, confirmed by the vowel space (i.e., convex hull) areas for the casual and Clear speech: 2.32 and 3.93 ($x10^5 Hz^2$), respectively. Moreover, the vowel space shape is largely maintained in this expansion, assuring that the perceptual distinctions among vowels are respected. Additionally, Fig.1b shows the F1 and F2 differences between the Clear and casual vowel spaces, i.e., the amount that each formant in the casual vowel space needs to be shifted in order to match the corresponding formant in the Clear vowel space. Rather than a uniform increase or decrease, it can be seen that both F1 and F2 are shifted either up or down, depending on the formant frequency. Specifically, low F1 or F2 of the casual speech are decreased, while high F1 or F2 are increased going from casual to Clear, ultimately expanding the vowel space. This trend is shown more explicitly by fitting the formant shifts with a linear regression, as shown by the solid lines in Fig.1b. Then, in order to visualize frequency shifts of a piecewise linear function that would emulate an expanding vowel space, a linear interpolation between the F1 and F2 boundaries and a return to zero-shift at the endpoints is also indicated. Ultimately, the overall form of this function inspires the frequency warping approach considered in this work.



(a) (b)

Figure 1: (a) Casual and Clear vowel spaces. (b) Casual-to-Clear formant shifts with piecewise linear fitting.

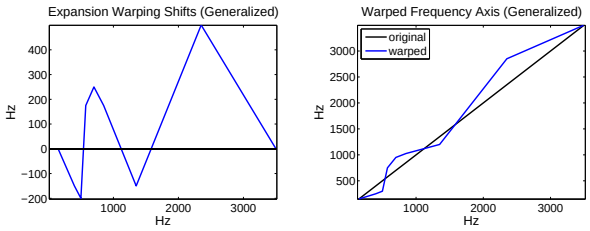
3. Frequency Warping for V.S. Expansion

3.1. Method Description

The proposed frequency warping approach for vowel space expansion can be described in two stages. The first defines a curve of generalized warping shifts, inspired by the formant shifts observed in the Clear speech vowel expansion. The second stage then outlines the frequency warping algorithm based on sampling the aforementioned curve on a frame-by-frame basis. The following respectively describes these stages in more detail.

3.1.1. *Clear Speech-Inspired V.S. Expansion Shifts*

Working from the trends shown in Fig.1b, the curve $\Delta(f)$ of generalized warping shifts (cf Fig.2a) used in the proposed approach was determined after several trials observing warped vowel spaces and by taking into account certain practical considerations. First, it is noted from Fig.1 that the magnitude of the formant shifts, on average, is quite small, especially compared to the separation of harmonic peaks (e.g., about 150 Hz on average) in the amplitude spectrum. Consequently, to define the curve of generalized warping shifts, the magnitude of the shifts must be significantly larger in order to overcome the harmonic structure in the amplitude spectrum and effectively shift a formant. Second, since the span of F1 is less than F2, care should be taken such that the F1 warping will achieve the desired effect without approaching DC or overlapping with F2. Consequently, the slope from negative to positive F1 shifts is exaggerated and the maximal shift is rounded out. Finally, the defined shifts should ensure that any warped frequency axis is always monotonically non-decreasing. Fig.2b displays the warped frequency axis generated from $\Delta(f)$, confirming that the slope is always non-negative. Fig.2b similarly indicates bounds on all possible warped axes. Specifically, the warped frequency axis of any frame will be defined as a set of lines connecting points (i.e., detected spectral peak frequencies) on the generalized warped axis.



(a) (b)

Figure 2: (a) $\Delta(f)$ - Generalized curve of exaggerated warping shifts. (b) Corresponding warped frequency axis.

3.1.2. *Frequency Warping Algorithm*

Before outlining the proposed frequency warping approach, some details of the speech analysis and synthesis are presented. Specifically, the analysis and synthesis is pitch-asynchronous, using a 30msec Hamming window and 10msec step. Each frame is analyzed using a 2048-pt DFT. In the case of speech modification, frequency warping filters are applied to the amplitude spectrum of a frame before synthesis using the inverse DFT (with the original phases) and overlap-add.

Now, let the spectral envelope from frame n of unmodified speech, $S_n^X(f)$, be represented here by the True Envelope using a cepstral order of 48 [16]. The frequency warping filter for frame n, $H_n^W(f)$, is then defined as

$$H_n^W(f) = S_n^W(f)/S_n^X(f) \quad (1)$$

where $S_n^W(f)$ is the warped spectral envelope

$$S_n^W(f) = S_n^X(W_n^{-1}(f)) \quad (2)$$

and $W_n(f)$ is the warping function for frame n, defined as follows. First, the spectral tilt, $S_n^{Xtilt}(f)$, is generated from the first two cepstral coefficients (zeroth and first order) of the True Envelope analysis. The spectral envelope peaks in the warping frequency range $f \in [150Hz, 3500Hz]$ are then detected from the tilt-normalized spectral envelope as

$$f_{n,i}^X = peak\_detect(S_n^X(f)/S_n^{Xtilt}(f)) \quad (3)$$

where $f_{n,i}^X$ indicates the frequency of the i^{th} spectral peak detected in frame n, $i = 1, ...M_n$. The peak detection algorithm defines peaks as local maxima preceded by local minima that are more than 10% lower than the maximum value in the frequency range (so as to avoid ripples or slight fluctuations and inflection points in the envelope). Next, the detected peaks for frame n sample $\Delta(f)$ to provide the intervals defining the frequency warping for the frame. Note that using the detected spectral peaks in this way tailors the frequency warping to the acoustic characteristics of the frame, while avoiding explicit formant estimation (e.g., limiting estimated peaks to F1 and F2), which can be quite error-prone. Specifically, the warped spectral peak frequencies are given by

$$f_{n,i}^W = f_{n,i}^X + \Delta(f_{n,i}^X) \quad (4)$$

and these frequencies, together with $f_{n,i}^X$, define a piecewise linear warping function with the form given in [13], [14]. Specifically, for $f \in [f_{n,i}^X, f_{n,i+1}^X]$, the warping function for frame n is

$$W_n(f) = A_{n,i}f + B_{n,i} \quad (5)$$

where $f_{n,0}^X = f_{n,0}^W = 150Hz, f_{n,M_n+1}^X = f_{n,M_n+1}^W = 3500Hz$, and

$$A_{n,i} = \frac{f_{n,i+1}^W - f_{n,i}^W}{f_{n,i+1}^X - f_{n,i}^X} = \Delta(f_{n,i+1}^X) - \Delta(f_{n,i}^X) \quad (6)$$

$$B_{n,i} = f_{n,i}^W - A_{n,i}f_{n,i}^X \quad (7)$$

With $W_n(f)$ defined from above, the warping filter $H_n(f)$ is calculated and applied to the amplitude spectrum of each frame. Application of the warping to all frames ensures that vowels spaces are expanded, without need for speech segmentation and labelling, while the influence on voiced non-vowels and unvoiced parts of speech is perceptually negligible (as confirmed upon listening to numerous warped speech samples). Furthermore, it should be emphasized that the focus on overall average trends for the vowel space expansion makes the proposed algorithm speaker-independent and thus generalized.

3.2. Results on Vowel Spaces

The vowel space for the frequency-warped casual speech is shown in Fig.3. The warped vowel space is generated in the same way as the casual and Clear vowel spaces in Section 2, but with the data being the the warped casual sentences. Fig.3 shows that the casual speech vowel space is successfully expanded. The warped vowel space area (3.58 compared to 2.32 $x10^5 Hz^2$), also confirms this expansion emulating that of Clear speech. Moreover, the structure of the vowel space is largely maintained, ensuring that the perceptual distinctions between vowels is respected, with only the distance or discriminability between them being increased. It should be noted that the proposed frequency warping approach is a generalized approximation, rather than deterministic replication, of the vowel space expansion observed in Clear speech. Overall, however, the vowel space expansion is achieved and largely respects observations from Clear speech via the proposed frequency warping, without explicit vowel or formant identification.

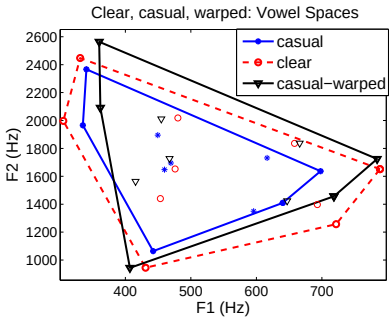


Figure 3: Clear, casual and casual-warped vowel spaces, with respective areas: 3.93, 2.32 and 3.58 ($x10^5 Hz^2$).

4. Evaluations of Intelligibility Impact

4.1. Extended Speech Intelligibility Index

Before describing listening tests, the following presents results of preliminary evaluations focusing on an objective measure, namely the extended Speech Intelligibility Index (extSII) [17]. The extSII was calculated using Speech Shaped Noise (SSN) added to yield a 0dB Signal to Noise ratio (SNR). Table 1 gives the average (median) of the extSII distributions for each of the conditions examined in evaluations. There are a few points to be gleaned from the extSII results. In terms of the frequency warping, there is a very slight gain observed in extSII over the unmodified casual speech, though this factor is probably too small to be meaningful. However, it should also be noted that the extSII fails to capture the intelligibility gain of Clear speech, highlighting some potential limitations of objective intelligibility metrics. Consequently, listening tests are required in an effort to capture more subtle acoustic modifications, such as vowel space expansion, in the clear and warped speech.

Table 1: Average extSII for Casual, Casual-Warped and Clear speech. The noise masker was SSN added to yield 0dB SNR.

	Casual	Casual-Warped	Clear
extSII	.311	.316	.312

4.2. Formal Listening Tests

In formal listening tests,[1] 20 native English-speakers heard LUCID sentences from each of the conditions (casual, casual-warped, Clear) described above, with SSN added at two levels to yield 0 and -4 dB SNR, respectively. It should be mentioned that, perceptually, no significant artefacts resulted from the frequency warping, though voice quality was noticeably altered. In the test, listeners were asked to type what they think they heard

[1]Thank you to Catherine Mayo and CSTR at the University of Edinburgh for their help administering the listening tests.

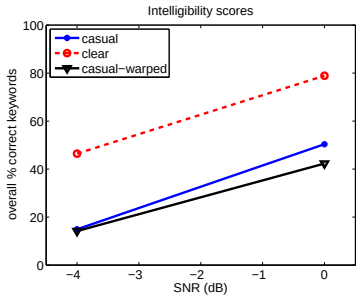


Figure 4: Intelligibility Test Scores for Casual, Casual-Warped and Clear speech. The percent of overall correct keyword identification is given for the low (-4dB) and high (0dB) SNR values with a SSN masker.

after hearing each sentence once. Of the listeners, 6 were removed due to inconsistencies in their scores, using established conditions (e.g. Clear speech) as a reference. Fig.4 shows the intelligibility scores (i.e., percent of words correctly identified) from the remaining listeners for each condition, at the high and low SNR levels. It should be noted that the inter-speaker variability was quite high and, in some cases, frequency warping did improve scores over unmodified casual speech. However, the general trend is displayed in Fig.4, indicating no significant improvement and even slight degradation overall.

In order to evaluate the statistical significance of these results, ANOVA tests were performed. Specifically, the ANOVA null hypothesis (i.e., the average values of the intelligibility scores for every condition are equal) was rejected using the F-test (SNR-4dB: $F(4,65)>28.719$, $p<0.05$ SNR0dB: $F(4,65)>25307$, $p<0.05$). Then, pairwise comparisons of the averages were performed using Fisher's Least Significant Difference (LSD) test in order to derive which of the groups differ significantly. The standardized difference between condition pairs is provided in Table 2. Analysis of the differences between the conditions confirms a significant categorical difference between Clear and casual speech, in agreement with observations from previous works. Similarly, the difference between the Clear and casual-warped speech is also significant. However, no significant categorical difference is found between the casual and casual-warped speech for either SNR. Thus, overall, there is essentially no intelligibility gain observed from the frequency warping for vowel space expansion.

Table 2: Results of Significant Difference Analysis between the Clear, casual and casual-warped intelligibility scores. The standardized differences are given and significant differences are indicated in bold.

SNR:	0dB (high)	-4dB (low)
Casual & Casual-Warped	1.41	.041
Casual & Clear	**5.14**	**5.66**
Casual-Warped & Clear	**6.55**	**5.70**

4.3. Discussion

The evaluation results can be explained from two perspectives. First, the lack of effectiveness of the frequency warping at increasing intelligibility could be due to the specific algorithm itself. That is, while successfully achieving vowel space expansion and altering voice characteristics without noticeable artefacts, the warping (by design) does not exactly replicate the expansion observed in Clear speech. As the authors do not claim that frequency warping or the present implementation is the only solution, perhaps there exists a more effective approach to expanding vowel space. Second, considering Clear speech, the observed vowel space expansion might be reflecting more important acoustic-phonetic modifications that occur on an increasingly detailed, spectro-temporally localized level. In other words, the observed vowel space expansion represents a static, average view of the articulation in the style that might not be highlighting the most pertinent cues. For example, the formant dynamics could be playing a significant role in enhancing the speech intelligibility and examination of these features in future work could prove fruitful. Of course, the two perspectives discussed above are inter-related. That is, in further identifying and localizing the pertinent acoustic-phonetic cues that lead to the vowel space expansion and also positively impact intelligibility, an appropriate modification approach to reproduce them could then be adopted.

5. Conclusions

This work presented and evaluated a novel approach to expand vowel space via frequency warping inspired by Clear speech analyses. Results indicate that, while vowel space is successfully expanded, there is no significant intelligibility gain from the frequency warping. These observations suggest that further detailed evaluation of the Clear speech intelligibility gain related to vowel space expansion is merited. In particular, a more localized level of analyses and consequent modifications (e.g. within phones and considering formant transitions) might expose more perceptually relevant differences that positively impact intelligibility.

6. Acknowledgements

This work was supported by LISTA. The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 25623.

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] J. Krause and L. Braida, "Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility." *JASA*, vol. 112(5), pp. 2165–2172, 2004.

[3] 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.

[4] 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.

[5] C. Davis and J. Kim, "Is speech produced in noise more

distinct and/or consistent?" in *Speech Science and Technology*, 2012, pp. 46–49.

[6] A. Kain, A. Amano-Kusumoto, and J. Hosom, "Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility," *J. Acous. Soc. Am.*, vol. 124, no. 4, pp. 2308–2319, 2008.

[7] R. Uchanski, S. C. SS, L. Braida, C. Reed, and N. Durlach, "Speaking clearly for the hard of hearing iv: Further studies of the role of speaking rate," *J Speech Hear Res.*, vol. 39, no. 3, pp. 494–509, 1996.

[8] M. Koutsogiannaki, M. Pettinato, C. Mayo, V. Kandia, and Y. Stylianou, "Can modified casual speech reach the intelligibility of clear speech?" in *Interspeech*, Portland Oregon, USA, 2012.

[9] I. McLoughlin and R. J. Chance, "Lsp-based speech modifications for intelligibility enhancement," in *13th Int. Conf. DSP*, vol. 2, 1997, pp. 591–594.

[10] 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.

[11] S. Mohammadi, A. Kain, and J. van Santen, "Making conversational vowels more clear," in *Interspeech*, Portland, Oregon, USA, 2012.

[12] H. Valbret, E. Moulines, and J. Tubach, "Voice transformation using psola technique," *SpeechComm*, vol. 11, no. 2-3, pp. 175 – 187, 1992.

[13] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Trans Audio, Speech, Lang Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[14] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans Audio, Speech, Lang Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[15] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2010. [Online]. Available: http://www.praat.org/

[16] A. Roebel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Digital Audio Effects (DAFx)*, 2005, pp. 30–35.

[17] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise." *J. Acous. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, 2006.