



**ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ**

Ψηφιακή Επεξεργασία Φωνής

Διάλεξη: Προσέγγιση Καθαρής Ομιλίας από
Τροποποιημένη Casual Ομιλία

Παρουσίαση: Μαρία Κουτσογιαννάκη

Στυλιανού Ιωάννης
Τμήμα Επιστήμης Υπολογιστών

Can modified casual speech reach the intelligibility of clear speech?

M. Koutsogiannaki

HY-578

November, 2012

1 Introduction

2 Modification techniques

- Examining the role of speaking rate in intelligibility and performing spectral modifications
- Refined time-scaling techniques with pause insertion
- Vowel space transformations using frequency warping

3 Conclusions

Outline

1 Introduction

2 Modification techniques

- Examining the role of speaking rate in intelligibility and performing spectral modifications
- Refined time-scaling techniques with pause insertion
- Vowel space transformations using frequency warping

3 Conclusions

Definition of clear speech

Clear speech is a speaking style adopted by talkers when speaking in difficult communication situations.

- Lombard speech (noisy environment)
- L2 oriented or hearing-impaired oriented speech
- infant or computer oriented speech

According to the speech style talkers adopt, different acoustic features are met.

Clear speech is a speaking style adopted by speakers in an attempt to maximize the clarity of their speech in non-noisy conditions and is proven to be more intelligible than

Casual speech.

How can we transform casual speech in clear speech?

The simplest algorithm

- Find a database of clear and casual speech and test that clear speech is indeed clear
- Find which acoustic features are different
- Transform the features of casual speech to mimic the features of clear speech
- Test the modified corpus under noise to check the intelligibility

Acoustic features

Clear and casual speech differ in

- f_0 value and range
- vowel space
- Long Term Amplitude Spectra between 1000-3100Hz (Lombard like) and above 4000Hz
- word and sentence duration (word elongation and pause insertion)

These features are not consistent to all speakers...

The impact of these features to intelligibility is not known.

Outline

1 Introduction

2 Modification techniques

- Examining the role of speaking rate in intelligibility and performing spectral modifications
- Refined time-scaling techniques with pause insertion
- Vowel space transformations using frequency warping

3 Conclusions

Does slowing down help?

- Examining the role of speaking rate in the clear's speech intelligibility
- Performing modifications on casual signals to enhance intelligibility
 - time-scaling
 - spectral modifications using Spectral Shaping and Dynamic Range Compression

Proposed method for evaluating the effect of duration and spectral modifications on speech intelligibility.

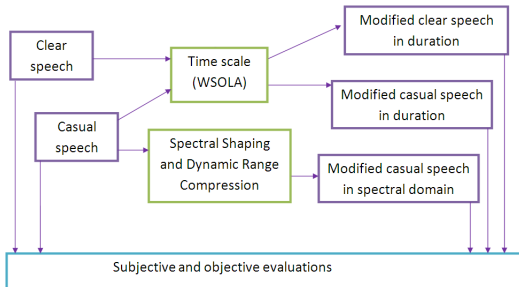


Figure 1: Methodology of evaluating the effect of duration and spectral modifications on speech intelligibility. Clear and casual speech sentences derive from the read speech of LUCID database

Spectral Shaping and Dynamic Range Compression

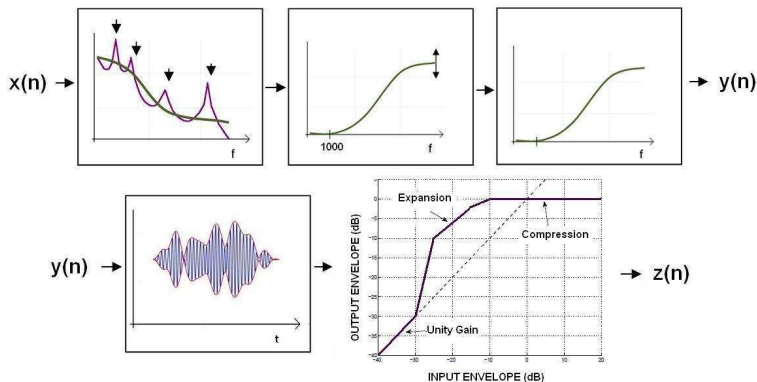


Figure 2: Performing Spectral Shaping and Dynamic Range Compression on signal $x(t)$

Subjective evaluations

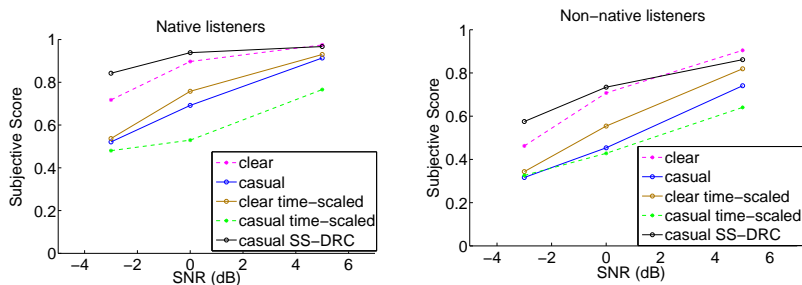


Figure 4: Subjective Measure Score for the 5 set of signals for different levels of SNR for Native (left) and Non-Native (right) Listeners

Objective evaluations

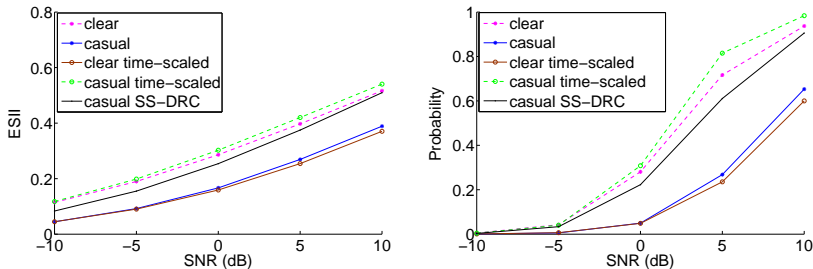


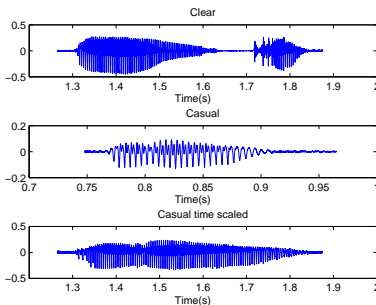
Figure 3: Objective Measure Score for the five sets of signals for different levels of SNR: Extended Speech Intelligibility Index (left) and Probability of correctly identifying a sentence (right)

Conclusions.

- clear speech is more intelligible than casual speech both for native and non-native speakers
- modified clear speech in higher speaking rates has lower intelligibility than unmodified clear speech
- modified clear speech in higher speaking rates has higher intelligibility than casual speech for mid and high SNRs
- modified casual speech by Spectral Shaping and Dynamic Range Compression has high intelligibility: **SS-DRC** modified casual speech gives greater intelligibility scores than clear speech in low and mid SNRs and similar intelligibility scores in high SNR. But... it is still not **clear!**

Conclusions..

- modified casual speech in lower speaking rates has lower intelligibility than unmodified casual speech (looking for an 'intelligent' time scale algorithm)



Conclusions...

- duration is not the only contributing factor to intelligibility. The time-expansion cannot fill the gap of the missing phonetic-level and acoustic-level information.
- motivation for 'smart' time scaling including reconstruction of missing parts and elegant pause insertions

Motivated...

Segmental time-alignment inappropriate for time-scaling

- lack of the naturalness
- absence of pauses and other acoustic information
- extreme elongation of some parts with “missing” cues

Proposed time-scaling methods

Respect naturalness of speech

- Uniform time-scaling

Mimic clear speech with some restrictions: add pauses and elongate stationary parts

- Perceptual Speech Quality measure model (PSQ)
- Rhythmogram inspired approach (RM)

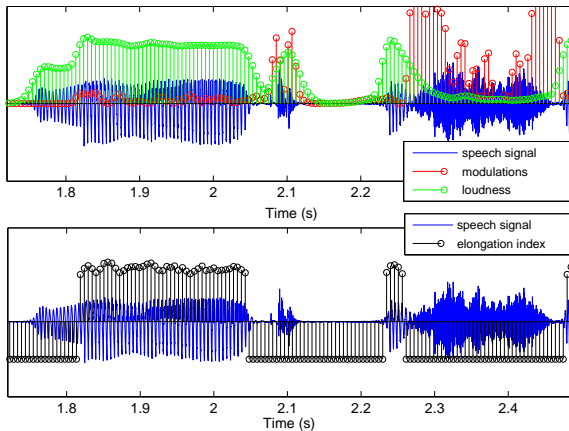
PSQ - elongation of stationary parts of speech

Detect stationary parts

- elongate stationary parts based on Elongation Index (EI), defined as:
$$EI = \begin{cases} L-M, & L-M < \text{threshold} \\ -1, & L-M > \text{threshold} \end{cases}$$
 - L: perceived loudness on the low frequency bands (0-300Hz)
 - M: loudness modulations of high frequency bands (around 4000Hz).
 - threshold values: [1.3 - 1.4].

Frames that have positive EI are time-scaled by WSOLA using a constant factor.

PSQ - example



Detection of non-stationary parts using PSQ model on the sentence "made a s(ign)" a) Loudness in low frequency bands and modulations in high frequency bands (top) b) Elongation index (bottom)

PSQ - detection and insertion of pauses

Detecting pauses

- normalized perceived loudness in the whole band.
- detection of valleys that are 30% lower than the maximum loudness
 - non-aggressive pauses: valleys with very low values, less than 10% of the normalized loudness
 - aggressive pauses: valleys that fall in the middle of word boundaries, (10%, 20%] of the normalized loudness

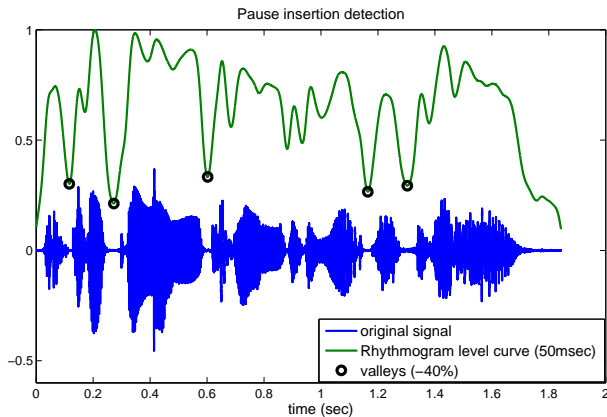
Inserting pauses

- adding a constant pause for the the non-aggressive
- preprocess the aggressive: time-scaling of the signal around the location where the gap will be inserted and apply a hamming window on the center of the valley

RM - pause detection and insertion

- rectifying speech signal and raising it to the one-third power
- convolving it to a “gross” Gaussian window (50msec length)
- detecting the deepest valleys of the resulting envelope
- inserting pauses proportional to the envelope valley depth

RM - example 1

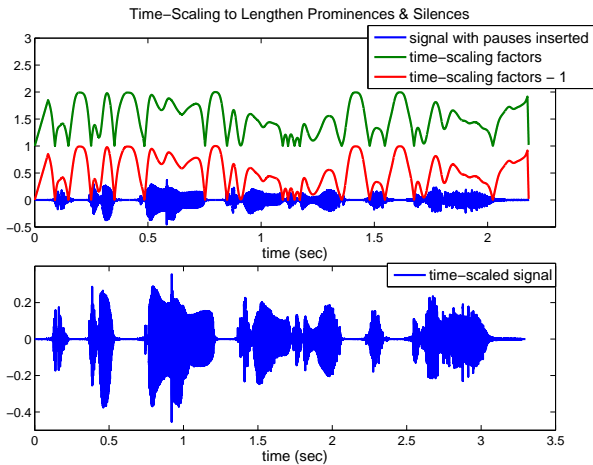


Rhythmogram-based pause detection.

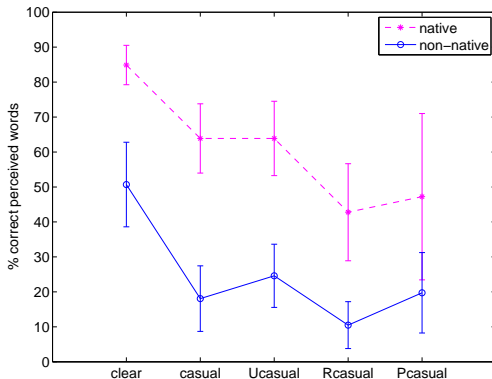
RM - time-scaling

- extract envelope curve for the “new” signal
- remove the mean of the normalized envelope
- rectify the envelope: valleys become peaks
 - non-stationary parts lie near zero
 - pauses will be time-scaled
- the rectified envelope, scaled by a maximum scaling factor+1 is the input to WSOLA

RM - example 2

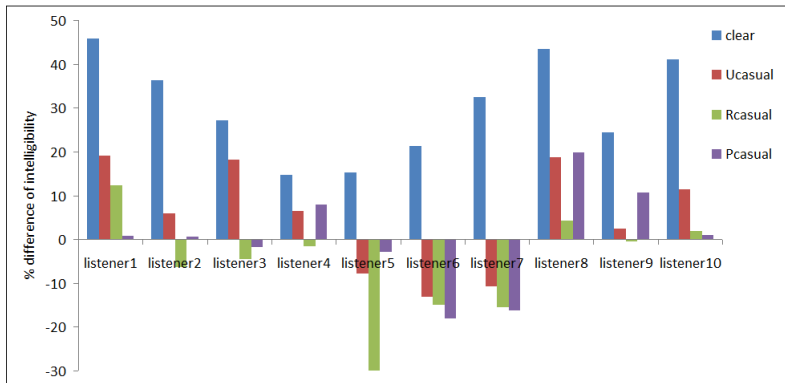


Evaluations 1/3



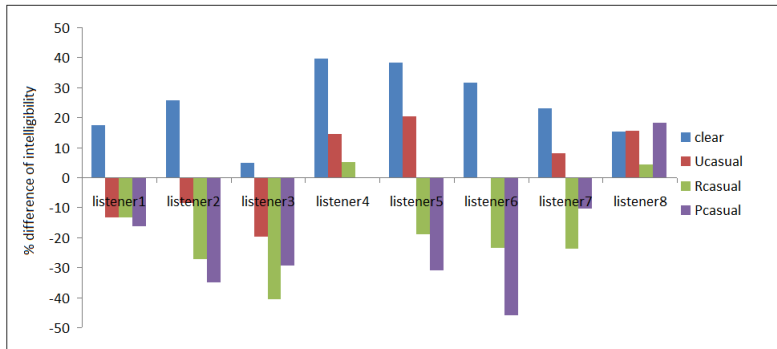
Subjective Intelligibility Score for the 5 set of signals for 0dB SNR. The percentage of correctly perceived words for each set for native and non-native listeners and the corresponding standard deviations. The Ucasual, Rcasual and Pcasual refer to the casual speech modified by the corresponding time-scaling techniques Uniform, Rhythmogram-based and PSQ-based.

Evaluations 2/3



Difference of the percentages of correctly perceived words between each set and the casual speech, for non-native. The Ucasual, Rcasual and Pcasual refer to the casual speech modified by the corresponding time-scaling techniques Uniform, Rhythmogram-based and PSQ-based.

Evaluations 3/3



Difference of the percentages of correctly perceived words between each set and the casual speech native listeners. The **Ucasual**, **Rcasual** and **Pcasual** refer to the casual speech modified by the corresponding time-scaling techniques Uniform, Rhythmogram-based and PSQ-based.

Conclusions

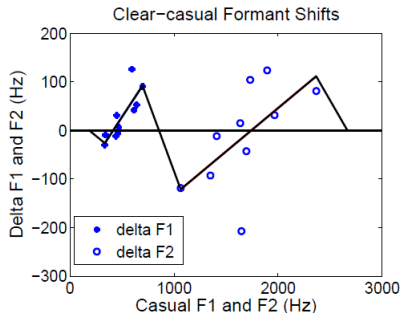
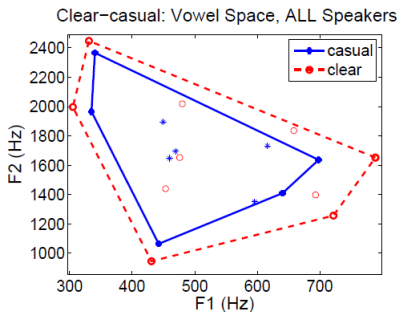
Proposed time-scaling techniques

- respect naturalness
- create artificial boundaries by successfully inserting pauses
- elongate speech without introducing artifacts

However,

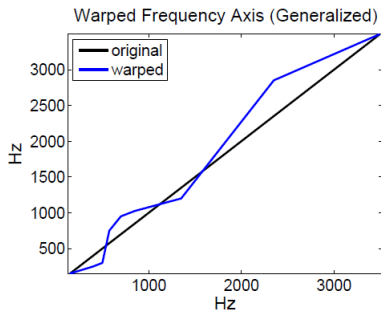
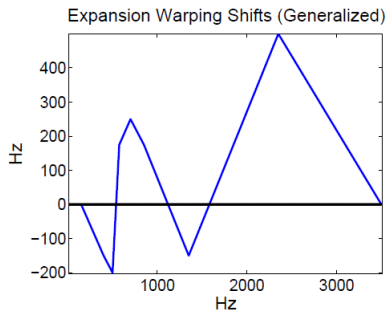
- PSQ and HM could not increase the intelligibility of casual speech:
 - naturalness is altered by elongating and inserting pauses
 - 61% of the speakers reported an intelligibility increase of modified casual speech with the uniform-time scaling
- the uniformly time-scaled casual did not sufficiently and/or consistently increase the intelligibility of casual speech

Vowel space differences on clear and casual speech



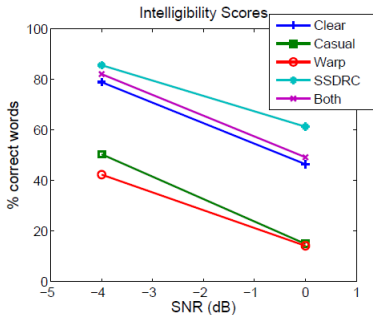
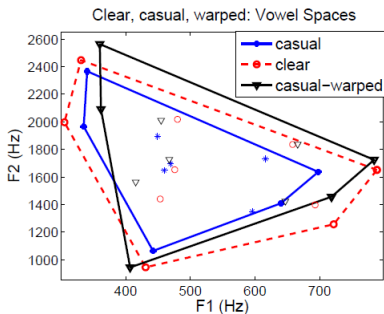
a) Vowel space of clear and casual speech. b) F1-F2 shifts

Defining a frequency warping function



a) F1-F2 exaggerated shifts b) Frequency warping function

Transforming the vowel space of casual speech - Results



a) **Casual and Casual warped** and clear vowel space b) Intelligibility tests

Outline

1 Introduction

2 Modification techniques





- Examining the role of speaking rate in intelligibility and performing spectral modifications
- Refined time-scaling techniques with pause insertion
- Vowel space transformations using frequency warping

3 Conclusions

Transforming casual speech to clear speech a tricky problem

- speaking rate helps but time-scaling “disturbs” naturalness
- vowel space expansion shows no increase of intelligibility
- spectral boosting helps in noise but the signal is no longer clear (in non-noisy conditions)
- f_0 doesn't seem to contribute to intelligibility

References

-  M. Koutsogiannaki, E. Godoy and Y. Stylianou
Towards Increasing Casual Speech Intelligibility with Time-Scaling and Pause Insertion Algorithms
Journal of Computer speech and language, submitted October 2012
-  M. Koutsogiannaki, M. Pettinato, C. Mayo, V. Kandia and Y. Stylianou
Can modified casual speech reach the intelligibility of clear speech?
Interspeech 2012
-  E. Godoy, M. Koutsogiannaki and Y. Stylianou
Acoustic Analyses of Lombard and Clear Speaking Styles: Towards Speech Modifications to Improve Intelligibility
Journal of Computer speech and language, submitted October 2012
-  E. Godoy, M. Koutsogiannaki and Y. Stylianou
Assessing intelligibility gains of vowel space expansion via Clear speech-inspired frequency warping
ICASSP 2013

References



M. Koutsogiannaki, C. Mayo, V. Kandia and Y. Stylianou
On the detection of intelligibility advantage of clear speech vs. casual speech
LISTA workshop 2012



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



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. 710, 2010.



Todd, N.P. and Brown, G.
Visualization of Rhythm Time and Meter
Artificial Intelligence Review, 1996

Τέλος Ενότητας



Ευρωπαϊκή Ένωση
Πρωτόκολλο Κοινωνίας Τεχνών



ΕΥΡΩΠΑΪΚΟ ΚΕΝΤΡΟ ΚΑΙΝΟΤΟΜΙΑΣ

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Κρήτης**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.



Σημειώματα

Σημείωμα αδειοδότησης

- Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά, Μη Εμπορική Χρήση, Όχι Παράγωγο Έργο 4.0 [1] ή μεταγενέστερη, Διεθνής Έκδοση. Εξαιρούνται τα αυτοτελή έργα τρίτων π.χ. φωτογραφίες, διαγράμματα κ.λ.π., τα οποία εμπεριέχονται σε αυτό και τα οποία αναφέρονται μαζί με τους όρους χρήσης τους στο «Σημείωμα Χρήσης Έργων Τρίτων».

[1] <http://creativecommons.org/licenses/by-nc-nd/4.0/>



- Ως **Μη Εμπορική** ορίζεται η χρήση:
 - που δεν περιλαμβάνει άμεσο ή έμμεσο οικονομικό όφελος από την χρήση του έργου, για το διανομέα του έργου και αδειοδόχο
 - που δεν περιλαμβάνει οικονομική συναλλαγή ως προϋπόθεση για τη χρήση ή πρόσβαση στο έργο
 - που δεν προσπορίζει στο διανομέα του έργου και αδειοδόχο έμμεσο οικονομικό όφελος (π.χ. διαφημίσεις) από την προβολή του έργου σε διαδικτυακό τόπο
- Ο δικαιούχος μπορεί να παρέχει στον αδειοδόχο ξεχωριστή άδεια να χρησιμοποιεί το έργο για εμπορική χρήση, εφόσον αυτό του ζητηθεί.

Σημείωμα Αναφοράς

Copyright Πανεπιστήμιο Κρήτης, Στυλιανού Ιωάννης. «Ψηφιακή Επεξεργασία Φωνής. Προσέγγιση Καθαρής Ομιλίας από Τροποποιημένη Casual Ομιλία».

Έκδοση: 1.0. Ηράκλειο/Ρέθυμνο 2015. Διαθέσιμο από τη δικτυακή διεύθυνση: <http://www.csd.uoc.gr/~hy578>