Speech Intelligibility

Yannis Stylianou

University of Crete, Computer Science Dept., Greece,

yannis@csd.uoc.gr

SPCC 2015

Communication barriers

- Detecting and understanding speech in noise plays a significant role in our communication with others
- Speech produced under background noise is not always intelligible ⇒ increase vocal effort when speaking to enhance the audibility of voice (Lombard effect)
- Conversational/casual speech is much less intelligible than clear speech for both normal-hearing (linguistically inexperienced listeners) and hearing-impaired listeners ⇒ try to speak more clearly

Communication barriers

- Detecting and understanding speech in noise plays a significant role in our communication with others
- Speech produced under background noise is not always intelligible $\Rightarrow$ increase vocal effort when speaking to enhance the audibility of voice (Lombard effect)
- Conversational/casual speech is much less intelligible than clear speech for both normal-hearing (linguistically inexperienced listeners) and hearing-impaired listeners $\Rightarrow$ try to speak more clearly

Communication barriers

- Detecting and understanding speech in noise plays a significant role in our communication with others
- Speech produced under background noise is not always intelligible $\Rightarrow$ increase vocal effort when speaking to enhance the audibility of voice (Lombard effect)
- Conversational/casual speech is much less intelligible than clear speech for both normal-hearing (linguistically inexperienced listeners) and hearing-impaired listeners $\Rightarrow$ try to speak more clearly

Observing Humans

- Lombard effect: higher energy in the mid-frequency region of the spectrum, reduced spectral tilt ...
- Clear speech: higher energy in the high-frequency region of the spectrum, expanded vowel space, slower speaking rate ...
- Nasals, onsets, offsets have low energy (speech production constraints)

Observing Humans

- Lombard effect: higher energy in the mid-frequency region of the spectrum, reduced spectral tilt ...
- Clear speech: higher energy in the high-frequency region of the spectrum, expanded vowel space, slower speaking rate ...
- Nasals, onsets, offsets have low energy (speech production constraints)

Observing Humans

- Lombard effect: higher energy in the mid-frequency region of the spectrum, reduced spectral tilt ...
- Clear speech: higher energy in the high-frequency region of the spectrum, expanded vowel space, slower speaking rate ...
- Nasals, onsets, offsets have low energy (speech production constraints)

Approaches to improve speech intelligibility

- High-pass filtering and amplitude compression (Niederjohn et al. 1976)
- Optimizing objective intelligibility criteria (e.g., SII, GP, STOI) (B. Sauert et al. 2006-2010, Y. Tang et al. 2012, R. Heusdens et al. 2012)
- Selective enhancement (V. Hazan et al. 1996, S.D.Yoo et al., 2007)

Approaches to improve speech intelligibility

- High-pass filtering and amplitude compression (Niederjohn et al. 1976)
- Optimizing objective intelligibility criteria (e.g., SII, GP, STOI) (B. Sauert et al. 2006-2010, Y. Tang et al. 2012, R. Heusdens et al. 2012)
- Selective enhancement (V. Hazan et al. 1996, S.D.Yoo et al., 2007)

Approaches to improve speech intelligibility

- High-pass filtering and amplitude compression (Niederjohn et al. 1976)
- Optimizing objective intelligibility criteria (e.g., SII, GP, STOI) (B. Sauert et al. 2006-2010, Y. Tang et al. 2012, R. Heusdens et al. 2012)
- Selective enhancement (V. Hazan et al. 1996, S.D.Yoo et al., 2007)

Spectral Shaping

- Probability of voicing: $P_v(t)$
- Adaptive spectral shaping:
  - Enhancement of spectral maxima:

$$H_s(\omega, t) = \left(\frac{E(\omega, t)}{T(\omega, t)}\right)^{\beta\ P_v(t)}$$

  - Pre-emphasis:

$$H_p(\omega, t) = \begin{cases} 1 & \omega \leq \omega_0 \\ 1 + \dfrac{\omega - \omega_0}{\pi - \omega_0} g\ P_v(t) & \omega > \omega_0 \end{cases}$$

- Fixed spectral shaping: $H_r(\omega)$ (boosting high frequencies)
- Spectral Shaping:

$$\hat{E}(\omega, t) = E(\omega, t)\ H_s(\omega, t) H_p(\omega, t) H_r(\omega)$$

Spectral Shaping

- Probability of voicing: $P_v(t)$
- Adaptive spectral shaping:
  - Enhancement of spectral maxima:

$$H_s(\omega,t) = \left(\frac{E(\omega,t)}{T(\omega,t)}\right)^{\beta\ P_v(t)}$$

  - Pre-emphasis:

$$H_p(\omega,t) = \begin{cases} 1 & \omega \le \omega_0 \\ 1 + \frac{\omega-\omega_0}{\pi-\omega_0} g\ P_v(t) & \omega > \omega_0 \end{cases}$$

- Fixed spectral shaping: $H_r(\omega)$ (boosting high frequencies)
- Spectral Shaping:

$$\hat{E}(\omega,t) = E(\omega,t)\ H_s(\omega,t) H_p(\omega,t) H_r(\omega)$$

Spectral Shaping

- Probability of voicing: $P_v(t)$
- Adaptive spectral shaping:
  - Enhancement of spectral maxima:

$$H_s(\omega, t) = \left(\frac{E(\omega, t)}{T(\omega, t)}\right)^{\beta \; P_v(t)}$$

  - Pre-emphasis:

$$H_p(\omega, t) = \begin{cases} 1 & \omega \leq \omega_0 \\ 1 + \frac{\omega - \omega_0}{\pi - \omega_0} g \; P_v(t) & \omega > \omega_0 \end{cases}$$

- Fixed spectral shaping: $H_r(\omega)$ (boosting high frequencies)
- Spectral Shaping:

$$\hat{E}(\omega, t) = E(\omega, t) \; H_s(\omega, t) H_p(\omega, t) H_r(\omega)$$

Spectral Shaping

- Probability of voicing: $P_v(t)$
- Adaptive spectral shaping:
  - Enhancement of spectral maxima:

$$H_s(\omega, t) = \left(\frac{E(\omega, t)}{T(\omega, t)}\right)^{\beta\ P_v(t)}$$

  - Pre-emphasis:

$$H_p(\omega, t) = \begin{cases} 1 & \omega \leq \omega_0 \\ 1 + \dfrac{\omega - \omega_0}{\pi - \omega_0} g\ P_v(t) & \omega > \omega_0 \end{cases}$$

- Fixed spectral shaping: $H_r(\omega)$ (boosting high frequencies)
- Spectral Shaping:

$$\hat{E}(\omega, t) = E(\omega, t)\ H_s(\omega, t) H_p(\omega, t) H_r(\omega)$$

Spectral Shaping

- Probability of voicing: $P_v(t)$
- Adaptive spectral shaping:
  - Enhancement of spectral maxima:

$$H_s(\omega, t) = \left(\frac{E(\omega, t)}{T(\omega, t)}\right)^{\beta \; P_v(t)}$$

  - Pre-emphasis:

$$H_p(\omega, t) = \begin{cases} 1 & \omega \leq \omega_0 \\ 1 + \dfrac{\omega - \omega_0}{\pi - \omega_0} g \; P_v(t) & \omega > \omega_0 \end{cases}$$

- Fixed spectral shaping: $H_r(\omega)$ (boosting high frequencies)
- Spectral Shaping:

$$\hat{E}(\omega, t) = E(\omega, t) \; H_s(\omega, t) H_p(\omega, t) H_r(\omega)$$

Spectral Shaping

- Probability of voicing: $P_v(t)$
- Adaptive spectral shaping:
  - Enhancement of spectral maxima:

$$H_s(\omega, t) = \left(\frac{E(\omega, t)}{T(\omega, t)}\right)^{\beta \ P_v(t)}$$

  - Pre-emphasis:

$$H_p(\omega, t) = \begin{cases} 1 & \omega \leq \omega_0 \\ 1 + \dfrac{\omega - \omega_0}{\pi - \omega_0} g \ P_v(t) & \omega > \omega_0 \end{cases}$$

- Fixed spectral shaping: $H_r(\omega)$ (boosting high frequencies)
- Spectral Shaping:

$$\hat{E}(\omega, t) = E(\omega, t) \ H_s(\omega, t) H_p(\omega, t) H_r(\omega)$$

Dynamic Range Compression (DRC)

- Speech envelope: analytic signal and moving average filtering
- Dynamic stage:

$$\hat{e}(n) = \begin{cases} a_r \hat{e}(n-1) + (1-a_r)e(n), & \text{if } e(n) < \hat{e}(n-1) \\ a_a \hat{e}(n-1) + (1-a_a)e(n), & \text{if } e(n) \geq \hat{e}(n-1) \end{cases}$$

- Static stage:

$$g(n) = 10^{(e_{out}(n) - e_{in}(n))/20}$$

  where $e_{in}(n) = 20\log_{10}(\hat{e}(n)/e_0)$, with e_0 being the reference level
- DRC: $s_g(n) = g(n)s(n)$

Dynamic Range Compression (DRC)

- Speech envelope: analytic signal and moving average filtering
- Dynamic stage:

$$\hat{e}(n) = \begin{cases} a_r \hat{e}(n-1) + (1-a_r)e(n), & \text{if } e(n) < \hat{e}(n-1) \\ a_a \hat{e}(n-1) + (1-a_a)e(n), & \text{if } e(n) \geq \hat{e}(n-1) \end{cases}$$

- Static stage:

$$g(n) = 10^{(e_{out}(n) - e_{in}(n))/20}$$

where $e_{in}(n) = 20\log_{10}(\hat{e}(n)/e_0)$, with e_0 being the reference level

- DRC: $s_g(n) = g(n)s(n)$

Dynamic Range Compression (DRC)

- Speech envelope: analytic signal and moving average filtering
- Dynamic stage:

$$\hat{e}(n) = \begin{cases} a_r \hat{e}(n-1) + (1-a_r)e(n), & \textit{if } e(n) < \hat{e}(n-1) \\ a_a \hat{e}(n-1) + (1-a_a)e(n), & \textit{if } e(n) \geq \hat{e}(n-1) \end{cases}$$

- Static stage:

$$g(n) = 10^{(e_{out}(n) - e_{in}(n))/20}$$

where $e_{in}(n) = 20 \log_{10}(\hat{e}(n)/e_0)$, with e_0 being the reference level

- DRC: $s_g(n) = g(n)s(n)$

DYNAMIC RANGE COMPRESSION (DRC)

- Speech envelope: analytic signal and moving average filtering
- Dynamic stage:

$$\hat{e}(n) = \begin{cases} a_r \hat{e}(n-1) + (1-a_r)e(n), & \text{if } e(n) < \hat{e}(n-1) \\ a_a \hat{e}(n-1) + (1-a_a)e(n), & \text{if } e(n) \geq \hat{e}(n-1) \end{cases}$$

- Static stage:

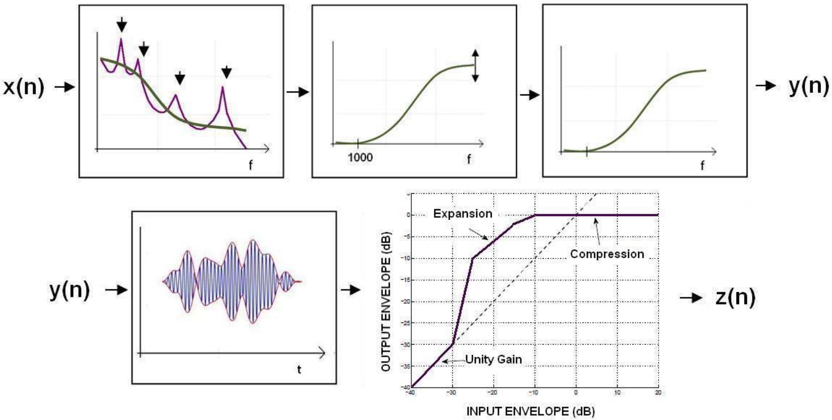
$$g(n) = 10^{(e_{out}(n)-e_{in}(n))/20}$$

  where $e_{in}(n) = 20\log_{10}(\hat{e}(n)/e_0)$, with e_0 being the reference level

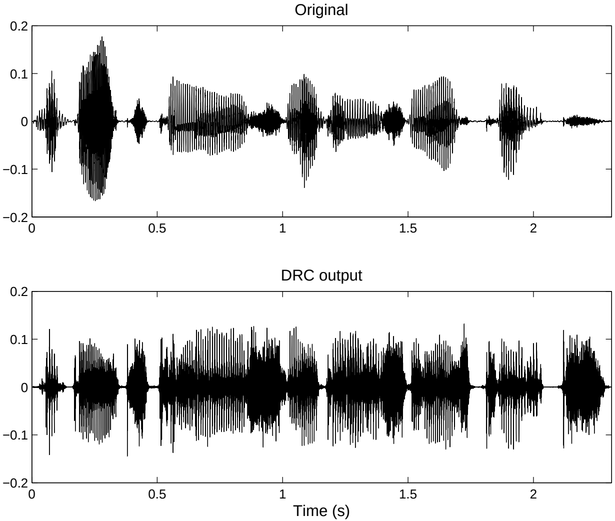
- DRC: $s_g(n) = g(n)s(n)$

SSDRC

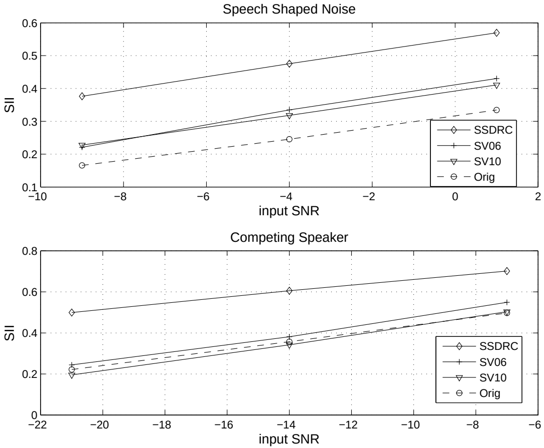
- Spectral Shaping and Dynamic Range Compression



SSDRC: Example of application



Objective Evaluation



- SV06: Sauert et al. 2006, SV10: Sauert et al. 2010

Formal Listening Test - Hurricane Challenge

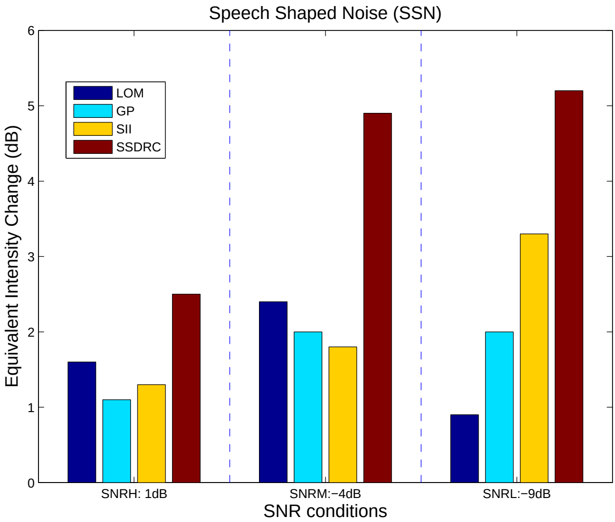
- 139 listeners whose native language was English
- Listeners received an audiological screening
- 6 conditions: 2 masker types $\times$ 3 SNR levels.
- 18 Harvard sets was mixed with noise for each of the 6 conditions
- We made sure that: each listener heard one block in each of the 18 noise conditions, no listener heard the same sentence twice, and each condition was heard by the same number of listeners.
- Each listener heard 180 sentences (apart from practice sentences)

Formal Listening Test:

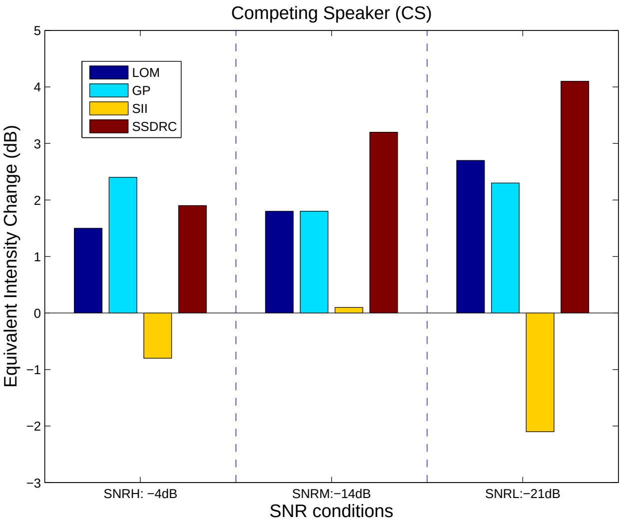
We compare:

- Normal speech
- Lombard speech [LOM]
- Spectral Modification optimizing GP (Y. Tang et al. 2012) [GP]
- Spectral Modification optimizing SII (B. Sauert et al. 2011) [SII]
- Suggested approach [SSDRC]

Formal Listening Test: SSN



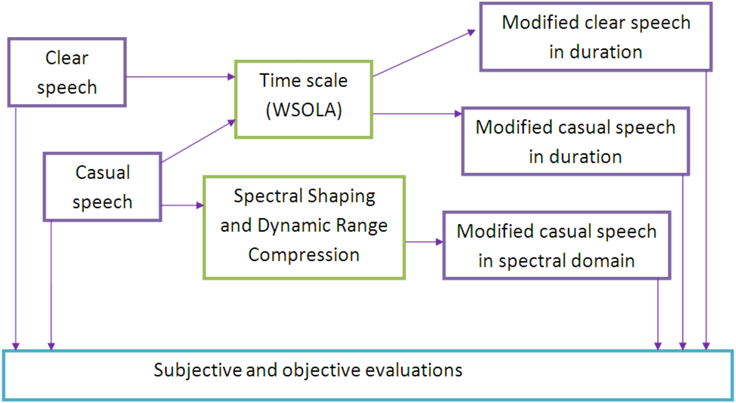
Formal Listening Test: CS



Corpus of clear and casual speech signals

- Read speech from the LUCID database: read speech is an exaggerated form of clear speech relative to the spontaneous clear speech (V. Hazan and R. Baker, 2010)
  - Southern British English speakers producing clear and casual speech
  - meaningful sentences simple in syntax
  - 70 distinct sentences are selected, uttered by 14 female speakers and 9 male speakers.

Evaluation procedure

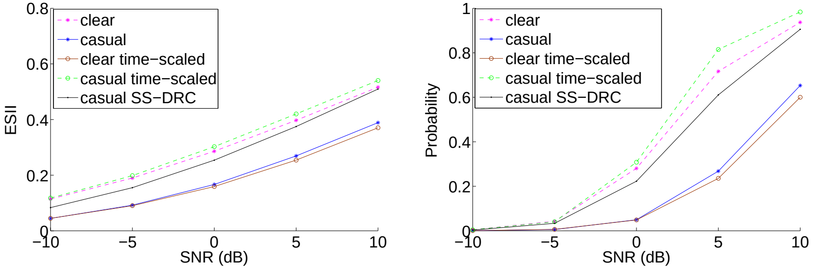


Objective and subjective evaluations

- Objective evaluations based on Extended Speech Intelligibility Index (ESII)
- Subjective evaluations: listening tests on duration and spectral modifications
- Speech Shaped Noise (SSN)

Objective evaluations

▷ Extended Speech Intelligibility Index (left) and Probability of correctly identifying a sentence (right)

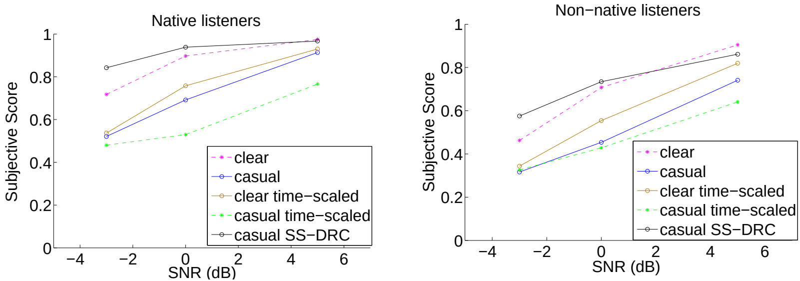


Subjective evaluations

- 70 different sentences
- five sets of signals at 3 different SNRs $\{-3, 0, 5\}$ dB
- 24 native and 15 non-native listeners
- scores from 1-5 according to intelligibility

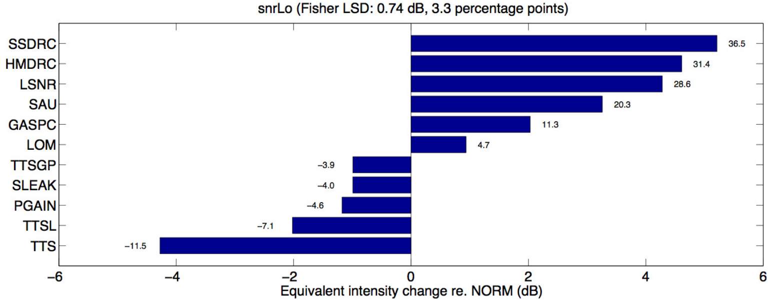
Subjective evaluations: Results

▷ Native (left) and Non-Native (right) Listeners



Motivation for using SSDRC

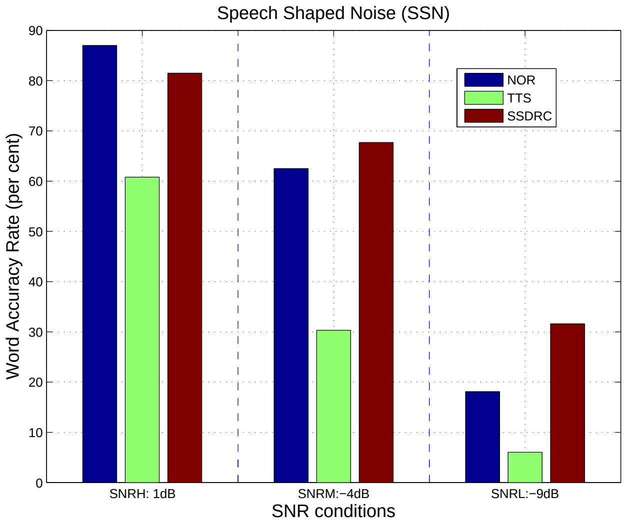
- SSN at -9dB SNR, N = 139 listeners



Formal Listening Test: synthetic speech

- 88 listeners whose native language was English
- Noise: 2 masker types $\times$ 3 SNR levels.
- 180 sentences were mixed with noise for each of the 6 conditions
- Each listener heard 180 sentences.
- No listener heard the same sentence twice.

Results: Synthetic Speech



Conclusions (1/3)

- Objectively and subjectively, SSDRC outperforms previous approaches and increases speech intelligibility in noise conditions
- For natural speech, SSDRC may provide up to 5 *dB* improvement in terms of Equivalent Intensity Change (EIC)
- For synthetic speech, SSDRC clearly increases its intelligibility:
  - in high SNR conditions, the intelligibility of natural speech is attained
  - for lower SNR conditions, the intelligibility of natural speech is exceeded ($> 30\%$)

Conclusions (1/3)

- Objectively and subjectively, SSDRC outperforms previous approaches and increases speech intelligibility in noise conditions
- For natural speech, SSDRC may provide up to 5 *dB* improvement in terms of Equivalent Intensity Change (EIC)
- For synthetic speech, SSDRC clearly increases its intelligibility:
  - in high SNR conditions, the intelligibility of natural speech is attained
  - for lower SNR conditions, the intelligibility of natural speech is exceeded ($> 30\%$)

Conclusions (1/3)

- Objectively and subjectively, SSDRC outperforms previous approaches and increases speech intelligibility in noise conditions
- For natural speech, SSDRC may provide up to 5 *dB* improvement in terms of Equivalent Intensity Change (EIC)
- For synthetic speech, SSDRC clearly increases its intelligibility:
  - in high SNR conditions, the intelligibility of natural speech is attained
  - for lower SNR conditions, the intelligibility of natural speech is exceeded ($>$ 30%)

Conclusions (1/3)

- Objectively and subjectively, SSDRC outperforms previous approaches and increases speech intelligibility in noise conditions
- For natural speech, SSDRC may provide up to 5 *dB* improvement in terms of Equivalent Intensity Change (EIC)
- For synthetic speech, SSDRC clearly increases its intelligibility:
  - in high SNR conditions, the intelligibility of natural speech is attained
  - for lower SNR conditions, the intelligibility of natural speech is exceeded ($>$ 30%)

Conclusions (1/3)

- Objectively and subjectively, SSDRC outperforms previous approaches and increases speech intelligibility in noise conditions
- For natural speech, SSDRC may provide up to 5 *dB* improvement in terms of Equivalent Intensity Change (EIC)
- For synthetic speech, SSDRC clearly increases its intelligibility:
  - in high SNR conditions, the intelligibility of natural speech is attained
  - for lower SNR conditions, the intelligibility of natural speech is exceeded ($> 30\%$)

Conclusions (2/3)

- 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 SSDRC has high intelligibility: SSDRC modified casual speech gives greater intelligibility scores than clear speech in low and mid SNRs and similar intelligibility scores in high SNR

Conclusions (2/3)

- 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 SSDRC has high intelligibility: SSDRC modified casual speech gives greater intelligibility scores than clear speech in low and mid SNRs and similar intelligibility scores in high SNR

Conclusions (2/3)

- 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 SSDRC has high intelligibility: SSDRC modified casual speech gives greater intelligibility scores than clear speech in low and mid SNRs and similar intelligibility scores in high SNR

Conclusions (2/3)

- 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 SSDRC has high intelligibility: SSDRC modified casual speech gives greater intelligibility scores than clear speech in low and mid SNRs and similar intelligibility scores in high SNR

Conclusions (3/3)

- SSDRC: Improves intelligibility of synthetic speech
- It can be applied as a post-filter, in the database, or in the parameter generation algorithm.
- In some cases, intelligibility of modified synthetic speech is higher than that of unmodified natural clear speech

Conclusions (3/3)

- SSDRC: Improves intelligibility of synthetic speech
- It can be applied as a post-filter, in the database, or in the parameter generation algorithm.
- In some cases, intelligibility of modified synthetic speech is higher than that of unmodified natural clear speech

Conclusions (3/3)

- SSDRC: Improves intelligibility of synthetic speech
- It can be applied as a post-filter, in the database, or in the parameter generation algorithm.
- In some cases, intelligibility of modified synthetic speech is higher than that of unmodified natural clear speech

Take home message

- SSDRC: Signal-processing based approach combining previous knowledge from speech-in-noise and clear/casual speaking styles literature
- Frame-based approach, no noise measurement ⇒ real time processing (real-time demo of SSDRC will be shown tomorrow afternoon)

Take home message

- SSDRC: Signal-processing based approach combining previous knowledge from speech-in-noise and clear/casual speaking styles literature
- Frame-based approach, no noise measurement ⇒ real time processing (real-time demo of SSDRC will be shown tomorrow afternoon)

Acknowledgment

- Varvara Kandia, Tudor-Cătălin Zorilă, Maria Koutsogiannaki, and Elizabeth Godoy for research on intelligibility
- Vassilis Tsiaras for the real-time version of SSDRC
- Martin Cooke, Cassia Valentini-Botinhao, Cassie Mayo, and Vassilis Karaiskos for organizing Hurricane Challenges
- EU-funded Future and Emerging Technologies (FET) project *The Listening Talker* (LISTA)

Thank you for your attention