CS578- Speech Signal Processing

Lecture on Intelligibility

Yannis Stylianou

University of Crete, Computer Science Dept., Multimedia Informatics Lab
yannis@csd.uoc.gr

Univ. of Crete

Outline

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 clear

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 clear

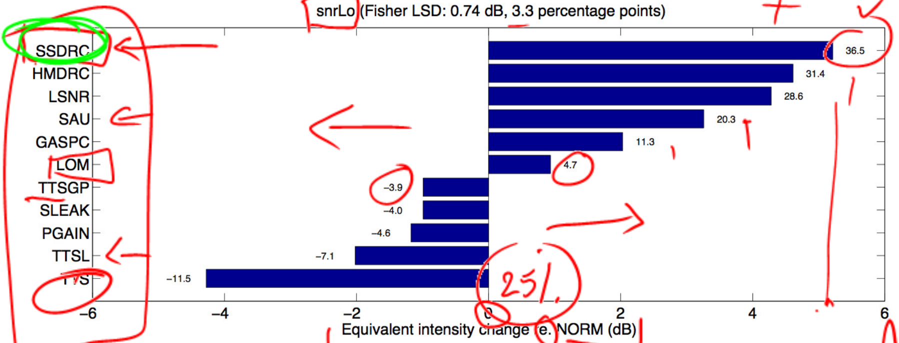
Communication barriers

- 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 clear

Motivation

Speech Shaped Noise

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



EIC

plain speech

Outline

LISTA, 2010-2013

- Current speech output technologies lack an essential element of human interaction, namely the ability to listen while talking
- Investigate how talkers react to changes in the listening environment,
- Apply this information to develop novel techniques for spoken output generation of artificial and natural speech.
- `http://listening-talker.org/`
- Hurricane Challenge

Speech material

- Phonetically-balanced sentences more representative of everyday speech
- Harvard sentence: "The key you designed will fit the lock"
- Male native English talker: 72 lists $\times 10$ sentences, very good recording conditions
- Post-processing: Downsampling to 16kHz, removing low-frequency artefacts, adding low amplitude (inaudible) random noise to the beginning and end of each sentence
- Hurricane Challenge: Only sets 1-18 (180 sentences) were used

Speech material

- Phonetically-balanced sentences more representative of everyday speech
- Harvard sentence: "The key you designed will fit the lock"
- Male native English talker: 72 lists $\times 10$ sentences, very good recording conditions
- Post-processing: Downsampling to 16kHz, removing low-frequency artefacts, adding low amplitude (inaudible) random noise to the beginning and end of each sentence
- Hurricane Challenge: Only sets 1-18 (180 sentences) were used

Speech material

- Phonetically-balanced sentences more representative of everyday speech
- Harvard sentence: "The key you designed will fit the lock"
- Male native English talker: 72 lists $\times 10$ sentences, very good recording conditions
- Post-processing: Downsampling to 16kHz, removing low-frequency artefacts, adding low amplitude (inaudible) random noise to the beginning and end of each sentence
- Hurricane Challenge: Only sets 1-18 (180 sentences) were used

Speech material

- Phonetically-balanced sentences more representative of everyday speech
- Harvard sentence: "The key you designed will fit the lock"
- Male native English talker: 72 lists $\times 10$ sentences, very good recording conditions
- Post-processing: Downsampling to 16kHz, removing low-frequency artefacts, adding low amplitude (inaudible) random noise to the beginning and end of each sentence
- Hurricane Challenge: Only sets 1-18 (180 sentences) were used

Speech material

- Phonetically-balanced sentences more representative of everyday speech
- Harvard sentence: "The key you designed will fit the lock"
- Male native English talker: 72 lists $\times 10$ sentences, very good recording conditions
- Post-processing: Downsampling to 16kHz, removing low-frequency artefacts, adding low amplitude (inaudible) random noise to the beginning and end of each sentence
- Hurricane Challenge: Only sets 1-18 (180 sentences) were used

Maskers

- *Fluctuating Masker:* Female ('Nina') competing speaker (CS); Read news speech, Harvard-like sentences
- *Steady-State Masker:* Speech-Shaped Noise (SSN); long-term average speech spectrum estimated by 'Nina'

Maskers

- *Fluctuating Masker:* Female ('Nina') competing speaker (CS); Read news speech, Harvard-like sentences
- *Steady-State Masker:* Speech-Shaped Noise (SSN); long-term average speech spectrum estimated by 'Nina'

Speech-Noise Mixtures

- Reduce probability listeners hearing the same background more than once
- Each masker fragment was 1 second longer than the sentence: 500 ms leading and lagging noise.
- Speech levels were scaled to produce a given SNR in the region where the speech was present.
- Intelligibility was evaluated at 3 SNRs for each masker type, expected to produce keyword scores of approximately 25, 50 and 75%.

Speech-Noise Mixtures

- Reduce probability listeners hearing the same background more than once
- Each masker fragment was 1 second longer than the sentence: 500 ms leading and lagging noise.
- Speech levels were scaled to produce a given SNR in the region where the speech was present.
- Intelligibility was evaluated at 3 SNRs for each masker type, expected to produce keyword scores of approximately 25, 50 and 75%.

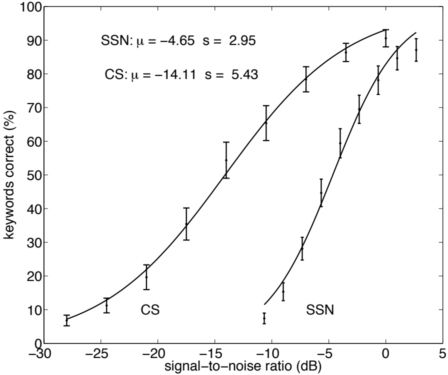
Speech-Noise Mixtures

- Reduce probability listeners hearing the same background more than once
- Each masker fragment was 1 second longer than the sentence: 500 ms leading and lagging noise.
- Speech levels were scaled to produce a given SNR in the region where the speech was present.
- Intelligibility was evaluated at 3 SNRs for each masker type, expected to produce keyword scores of approximately 25, 50 and 75%.

Speech-Noise Mixtures

- Reduce probability listeners hearing the same background more than once
- Each masker fragment was 1 second longer than the sentence: 500 ms leading and lagging noise.
- Speech levels were scaled to produce a given SNR in the region where the speech was present.
- Intelligibility was evaluated at 3 SNRs for each masker type, expected to produce keyword scores of approximately 25, 50 and 75%.

Baselines results



Two-parameter fitting logistic function:

$$p_n = \frac{1}{1 + e^{-(snr - a_n)/b_n}}$$

Equivalent Intensity Change (EIC)

- Inverse of logistic approximation to SNR-intelligibility function for speech style m and masker n:

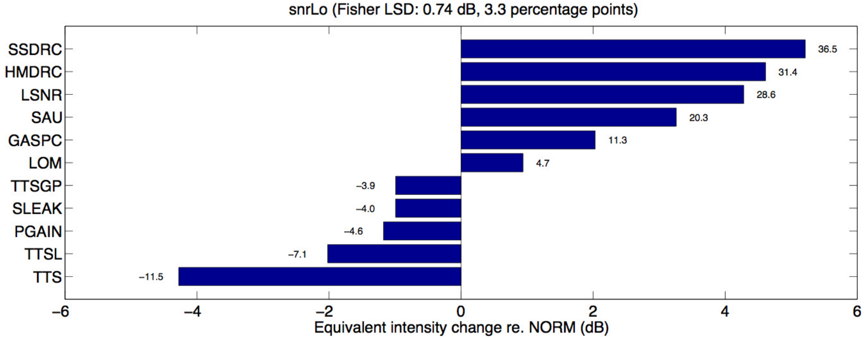
$$snr_{m,n} = a_n - b_n \log\left(\frac{1}{p_{m,n}} - 1\right)$$

- Equivalent Intensity Change (EIC):

$$EIC_{m,n} = snr_{m,n} - snr_{NORM}$$

Results

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



Outline

Approaches to improve speech intelligibility

- High-pass filtering and amplitude compression (Niederjohn et al. 1976 [1])
- Optimizing objective intelligibility criteria (e.g., SII, GP, STOI) (B. Sauert et al. 2006-2011 [2][3][4], Y. Tang et al. 2012 [5], C.H. Taal et al. 2012 [6])
- Selective enhancement (V. Hazan et al. 1996 [7], S.D.Yoo et al., 2007 [8])

Approaches to improve speech intelligibility

- High-pass filtering and amplitude compression (Niederjohn et al. 1976 [1])
- Optimizing objective intelligibility criteria (e.g., SII, GP, STOI) (B. Sauert et al. 2006-2011 [2][3][4], Y. Tang et al. 2012 [5], C.H. Taal et al. 2012 [6])
- Selective enhancement (V. Hazan et al. 1996 [7], S.D.Yoo et al., 2007 [8])

Approaches to improve speech intelligibility

- High-pass filtering and amplitude compression (Niederjohn et al. 1976 [1])
- Optimizing objective intelligibility criteria (e.g., SII, GP, STOI) (B. Sauert et al. 2006-2011 [2][3][4], Y. Tang et al. 2012 [5], C.H. Taal et al. 2012 [6])
- Selective enhancement (V. Hazan et al. 1996 [7], S.D.Yoo et al., 2007 [8])

Observations

- 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)

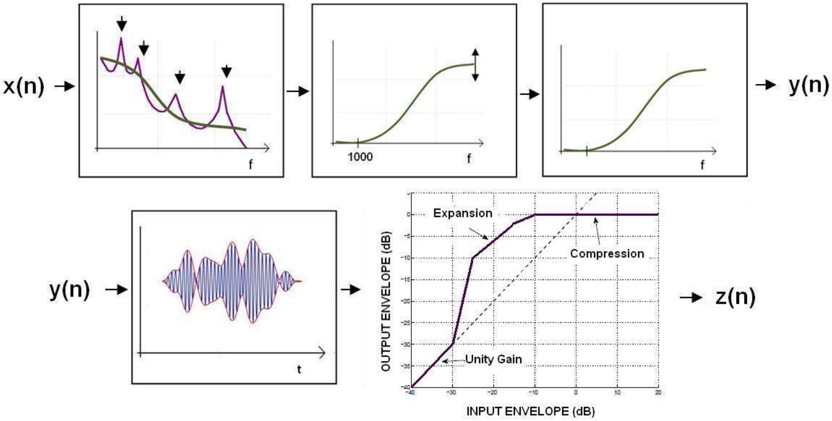
Observations

- 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)

Observations

- 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)

- Spectral Shaping and Dynamic Range Compression



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:

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

- 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:

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

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

Dynamic Range Compression (DRC)

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

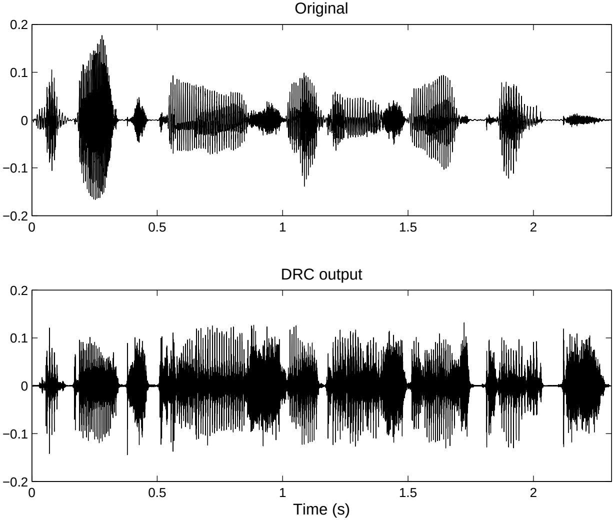
$$\hat{e}(n) = \begin{cases} a_r \hat{e}(n-1) + (1-a_r)e(n), & \textit{if } \ e(n) < \hat{e}(n-1) \\ a_a \hat{e}(n-1) + (1-a_a)e(n), & \textit{if } \ e(n) \geq \hat{e}(n-1) \end{cases}$$

- Static stage:

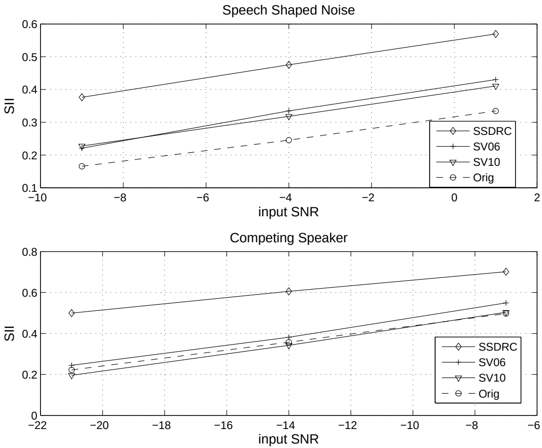
$$g(n) = 10^{(e_{out}(n) - e_{in}(n))/20}$$

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

SSDRC: Example of application



Objective Evaluation



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

Formal Listening Test - Hurricane Challenge

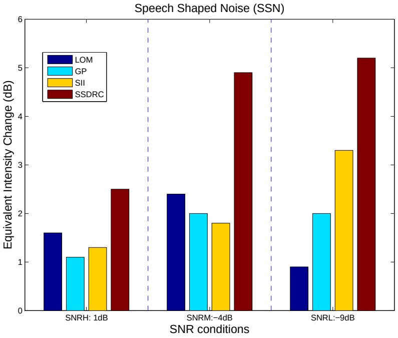
- 139 listeners whose native language was English
- Listeners received an audiological screening
- 6 conditions: 2 masker types $\times$ 3 SNR levels.
- 18 Harvard sets was mixed with noise for each of the 6 conditions
- We made sure that: each listener heard one block in each of the 18 noise conditions, 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)

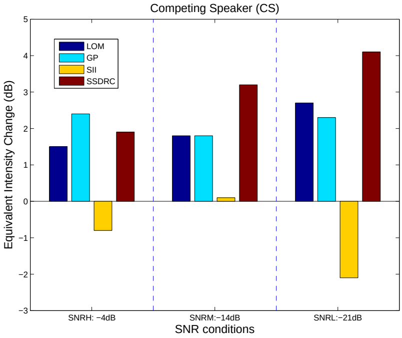
Formal Listening Test

We compare:

- Normal speech
- Lombard speech [LOM]
- Spectral Modification optimizing GP (Y. Tang et al. 2012) [GP][5]
- Spectral Modification optimizing SII (B. Sauert et al. 2011) [SII][9]
- Suggested approach (Zorila et al. 2012) [SSDRC] [10]

Formal Listening Test (near-field): SSN &CS

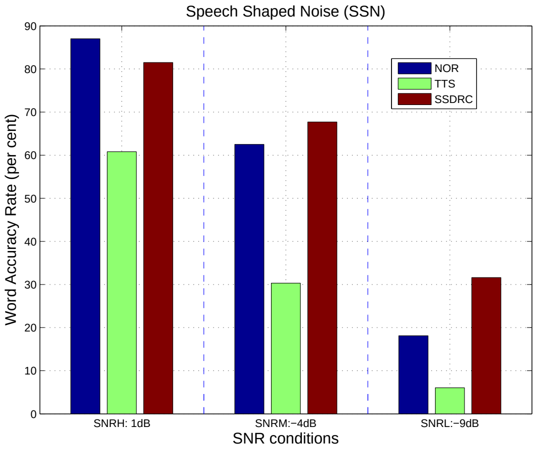




Formal Listening Test: synthetic speech

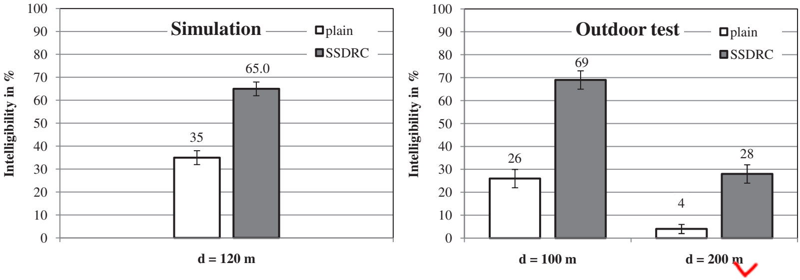
- 88 listeners whose native language was English
- Noise: 2 masker types $\times$ 3 SNR levels.
- 180 sentences were mixed with noise for each of the 6 conditions
- Each listener heard 180 sentences.
- No listener heard the same sentence twice.

Results (near-field): Synthetic Speech



- C. Valentini-Botinhao et al. *IS2013*[11]

Field Trial - far field



- T.C. Zorila, Y. Stylianou, T. Ishihara and M. Akamine: **Near and far field speech-in-noise intelligibility improvements based on a time-frequency energy reallocation approach** *IEEE, Trans. On Audio, Speech and Language Processing*, vol.24(10), Oct 2016, pp1808-1818

First conclusions

- SSDRC: Signal-processing based approach combining previous knowledge from speech-in-noise and clear/casual speaking styles literature
- Objectively and subjectively, SSDRC outperforms previous approaches
- 5 *dB* improvement in terms of Equivalent Intensity Change (EIC)
- Frame-based approach, no noise measurement $\Rightarrow$ real time processing
- Gains for near and far-field, various noise conditions

First conclusions

- SSDRC: Signal-processing based approach combining previous knowledge from speech-in-noise and clear/casual speaking styles literature
- Objectively and subjectively, SSDRC outperforms previous approaches
- 5 *dB* improvement in terms of Equivalent Intensity Change (EIC)
- Frame-based approach, no noise measurement $\Rightarrow$ real time processing
- Gains for near and far-field, various noise conditions

First conclusions

- SSDRC: Signal-processing based approach combining previous knowledge from speech-in-noise and clear/casual speaking styles literature
- Objectively and subjectively, SSDRC outperforms previous approaches
- 5 *dB* improvement in terms of Equivalent Intensity Change (EIC)
- Frame-based approach, no noise measurement $\Rightarrow$ real time processing
- Gains for near and far-field, various noise conditions

First conclusions

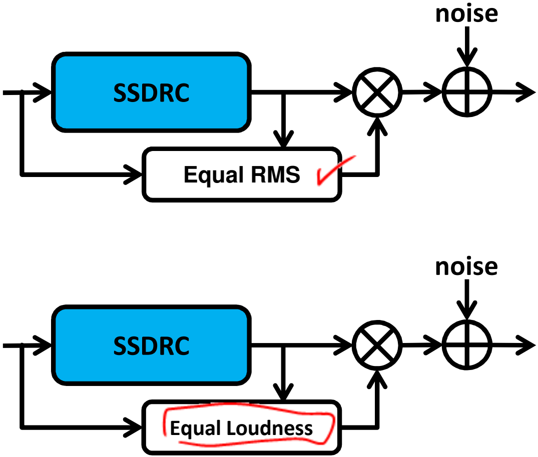
- SSDRC: Signal-processing based approach combining previous knowledge from speech-in-noise and clear/casual speaking styles literature
- Objectively and subjectively, SSDRC outperforms previous approaches
- 5 *dB* improvement in terms of Equivalent Intensity Change (EIC)
- Frame-based approach, no noise measurement $\Rightarrow$ real time processing
- Gains for near and far-field, various noise conditions

First conclusions

- SSDRC: Signal-processing based approach combining previous knowledge from speech-in-noise and clear/casual speaking styles literature
- Objectively and subjectively, SSDRC outperforms previous approaches
- 5 *dB* improvement in terms of Equivalent Intensity Change (EIC)
- Frame-based approach, no noise measurement $\Rightarrow$ real time processing
- Gains for near and far-field, various noise conditions

Outline

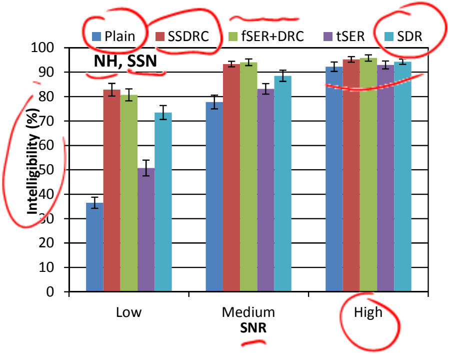
On Constraints

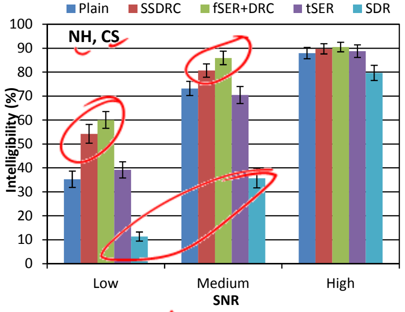


⇒ We need to repeat some experiments

- T.C. Zorila, Y. Stylianou, S. Flanagan and B.C.J. Moore: **Effectiveness of a loudness model for time-varying sounds in equating the loudness of sentences subjected to different forms of signal processing** *The Journal of the Acoustical Society of America*, vol.140(1), July 2016, pp1057-1061
- T.C. Zorila, Y. Stylianou, S. Flanagan and B.C.J. Moore: **Evaluation of Near-End Speech Enhancement under Equal-Loudness Constraint for Listeners with Normal-Hearing and Mild-to-Moderate Hearing Loss** *The Journal of the Acoustical Society of America*, vol.141(1), Jan 2017

Equal Loudness: Normal hearing



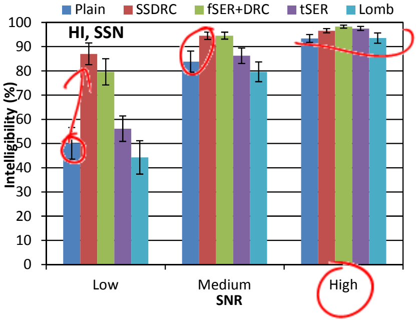


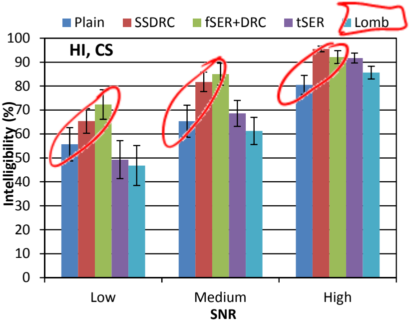
MIT

- **tSER:** time domain Spectral Energy Reallocation, Takou et al(IS2013)[12] based on Turiccia et al work (IEEE Trans 2005)[13]
- **fSER+DRC:** frequency domain Spectral Energy Reallocation and Dynamic Range Compression, Zorila et al. (IS2015)[14]

- **SDR:** Spectral Dynamic Recovery, Petko et al. (IEEE Trans 2015)[15]

Equal Loudness: Hearing impaired





- **tSER:** time domain Spectral Energy Reallocation, Takou et al(IS2013)[12] based on Turiccia et al work (IEEE Trans 2005)[13]

- **fSER+DRC:** frequency domain Spectral Energy Reallocation and Dynamic Range Compression, Zorila et al.

(IS2015)[14]

Further developments

- Clear-Casual Speech (E. Godoy et al., CSL 2014 [16])
- Synthetic Speech, (D. Erro et al. IEEE Trans 2014 [17])
- Special groups of listeners (S. Flanagan et al. Trends in Hearing 2018 [18])
- Noise-Dependent SSDRC (Griffin et al. ICASSP2015 [19])
- Special Session at IS2013 & Special Issue in Computer Speech and Language
- Real-time SSDRC (Show and Tell: IEEE ICASSP 2014 Florence, [20])

Further developments

- Clear-Casual Speech (E. Godoy et al., CSL 2014 [16])
- Synthetic Speech, (D. Erro et al. IEEE Trans 2014 [17])
- Special groups of listeners (S. Flanagan et al. Trends in Hearing 2018 [18])
- Noise-Dependent SSDRC (Griffin et al. ICASSP2015 [19])
- Special Session at IS2013 & Special Issue in Computer Speech and Language
- Real-time SSDRC (Show and Tell: IEEE ICASSP 2014 Florence, [20])

Further developments

- Clear-Casual Speech (E. Godoy et al., CSL 2014 [16])
- Synthetic Speech, (D. Erro et al. IEEE Trans 2014 [17])
- Special groups of listeners (S. Flanagan et al. Trends in Hearing 2018 [18])
- Noise-Dependent SSDRC (Griffin et al. ICASSP2015 [19])
- Special Session at IS2013 & Special Issue in Computer Speech and Language
- Real-time SSDRC (Show and Tell: IEEE ICASSP 2014 Florence, [20])

Further developments

- Clear-Casual Speech (E. Godoy et al., CSL 2014 [16])
- Synthetic Speech, (D. Erro et al. IEEE Trans 2014 [17])
- Special groups of listeners (S. Flanagan et al. Trends in Hearing 2018 [18])
- Noise-Dependent SSDRC (Griffin et al. ICASSP2015 [19])
- Special Session at IS2013 & Special Issue in Computer Speech and Language
- Real-time SSDRC (Show and Tell: IEEE ICASSP 2014 Florence, [20])

Further developments

- Clear-Casual Speech (E. Godoy et al., CSL 2014 [16])
- Synthetic Speech, (D. Erro et al. IEEE Trans 2014 [17])
- Special groups of listeners (S. Flanagan et al. Trends in Hearing 2018 [18])
- Noise-Dependent SSDRC (Griffin et al. ICASSP2015 [19])
- Special Session at IS2013 & Special Issue in Computer Speech and Language
- Real-time SSDRC (Show and Tell: IEEE ICASSP 2014 Florence, [20])

Further developments

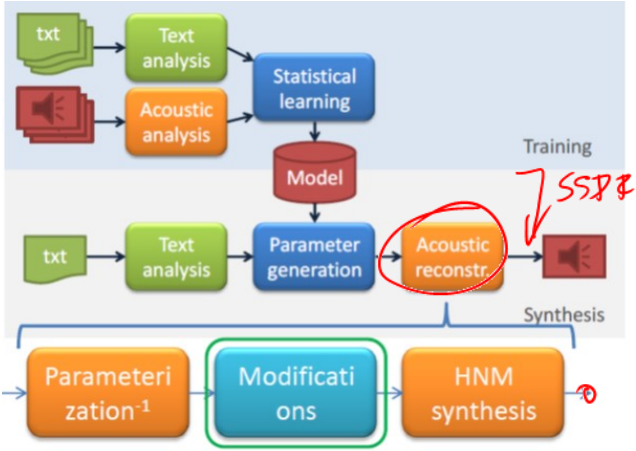
- Clear-Casual Speech (E. Godoy et al., CSL 2014 [16])
- Synthetic Speech, (D. Erro et al. IEEE Trans 2014 [17])
- Special groups of listeners (S. Flanagan et al. Trends in Hearing 2018 [18])
- Noise-Dependent SSDRC (Griffin et al. ICASSP2015 [19])
- Special Session at IS2013 & Special Issue in Computer Speech and Language
- Real-time SSDRC (Show and Tell: IEEE ICASSP 2014 Florence, [20])

Key papers to read

1. M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang: *Evaluating the intelligibility benefit of speech modifications in known noise conditions* Speech Communication, Jan 2013.
2. T.C. Zorila, V. Kandia, and Y. Stylianou: *Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression*, Interspeech 2012 !
3. M. Koutsogiannaki, M. Pettinato, C. Mayo, V. Kandia and Y. Stylianou: *Can modified casual speech reach the intelligibility of clear speech?*, Interspeech 2012
4. D. Erro, T.C. Zorila, Y. Stylianou, E. Navas and I. Hernaez: *Statistical Synthesizer with Embedded Prosodic and Spectral Modifications to Generate Highly Intelligible Speech in Noise*, Interspeech 2013
5. E. Godoy, C. Mayo, Y. Stylianou: *Increasing Speech Intelligibility via Spectral Shaping with Frequency Warping and Dynamic Range Compression plus Transient*, Interspeech 2013
6. C. Valentini-Botinhao, J. Yamagishi, S. King and Y. Stylianou: *Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of HMMbased synthetic speech in noise* Interspeech 2013

The issue for TTS

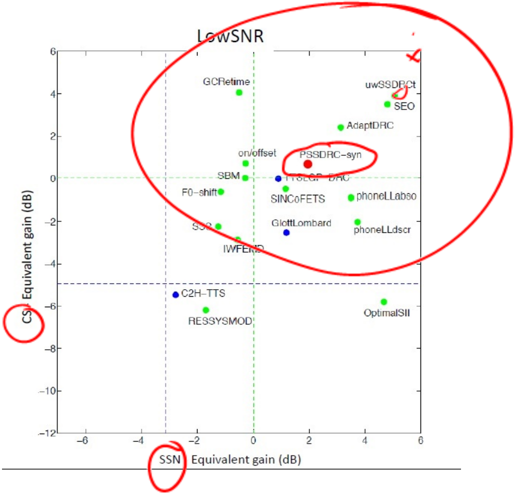
SSDRC LIKE POST-PROCESSING[17]



SSDRC

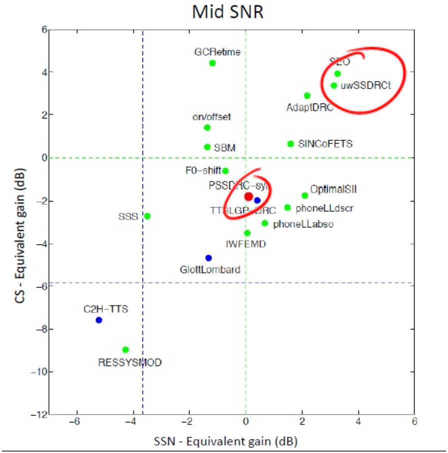
⇒ plus some modifications on duration and pitch

Hurricane II: Low SNRs



⇒ look for PSSDRC-syn

Hurricane II: Mid SNRs



⇒ look for PSSDRC-syn

Outline

ENRICH (2016-2019)

- **ENRICH:** Enriched communication across the lifespan. *MSCE, European Training Network*
- Transform speech to decrease its processing load, both universally and for individuals or populations of listeners
- Cognitive studies, modelling, engineering and real-world field evaluation with a range of listener groups
- Implementation of 14 projects, in three themes: 1) Reducing listening effort; 2) Enrichment and modalities; 3) Benefits for individuals and groups
- http://www.enrich-etn.eu/

ENRICH (2016-2019)

- **ENRICH:** Enriched communication across the lifespan. *MSCE, European Training Network*
- Transform speech to decrease its processing load, both universally and for individuals or populations of listeners
- Cognitive studies, modelling, engineering and real-world field evaluation with a range of listener groups
- Implementation of 14 projects, in three themes: 1) Reducing listening effort; 2) Enrichment and modalities; 3) Benefits for individuals and groups
- `http://www.enrich-etn.eu/`

ENRICH (2016-2019)

- **ENRICH:** Enriched communication across the lifespan.
  MSCE, European Training Network
- Transform speech to decrease its processing load, both universally and for individuals or populations of listeners
- Cognitive studies, modelling, engineering and real-world field evaluation with a range of listener groups
- Implementation of 14 projects, in three themes: 1) Reducing listening effort; 2) Enrichment and modalities; 3) Benefits for individuals and groups
- `http://www.enrich-etn.eu/`

ENRICH (2016-2019)

- **ENRICH:** Enriched communication across the lifespan. *MSCE, European Training Network*
- Transform speech to decrease its processing load, both universally and for individuals or populations of listeners
- Cognitive studies, modelling, engineering and real-world field evaluation with a range of listener groups
- Implementation of 14 projects, in three themes: 1) Reducing listening effort; 2) Enrichment and modalities; 3) Benefits for individuals and groups
- http://www.enrich-etn.eu/

ENRICH (2016-20~~19~~20)

- **ENRICH:** Enriched communication across the lifespan. *MSCE, European Training Network*
- Transform speech to decrease its processing load, both universally and for individuals or populations of listeners
- Cognitive studies, modelling, engineering and real-world field evaluation with a range of listener groups
- Implementation of 14 projects, in three themes: 1) Reducing listening effort; 2) Enrichment and modalities; 3) Benefits for individuals and groups
- `http://www.enrich-etn.eu/`

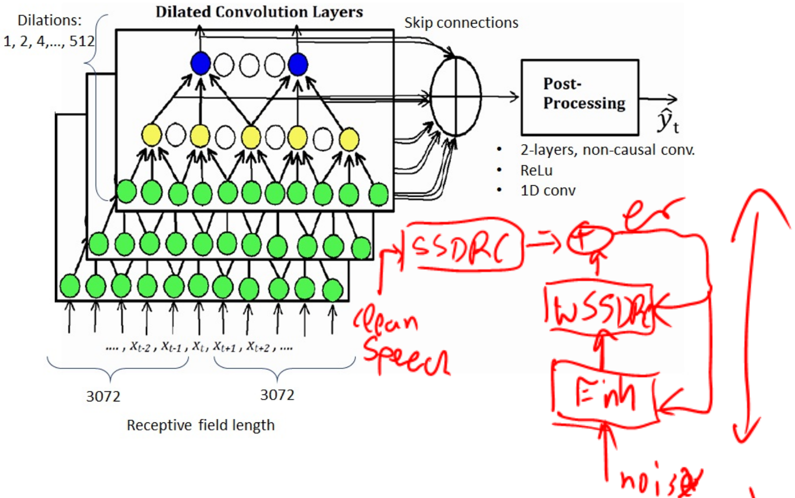
Focusing on two recent works

- **Wavenet-based SSDRC: wSSDRC:**
  Muhammed Shifas PV, Vassilis Tsiaras and Yannis Stylianou, *Speech intelligibility enhancement based on a non-causal Wavenet-like model*, Interpseech 2018, Hyderabad, India
- **Speaking style and listening effort:**
  Olympia Simantiraki, Martin Cooke, and Simon King, *Impact of different speech types on listening effort*, Interspeech 2018, Hyderabad, India

Focusing on two recent works

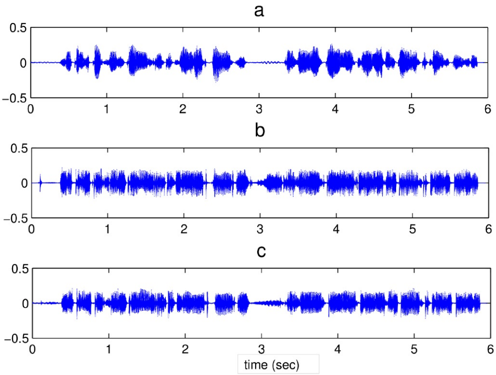
- **Wavenet-based SSDRC: wSSDRC:**
  Muhammed Shifas PV, Vassilis Tsiaras and Yannis Stylianou, *Speech intelligibility enhancement based on a non-causal Wavenet-like model*, Interpseech 2018, Hyderabad, India
- **Speaking style and listening effort:**
  Olympia Simantiraki, Martin Cooke, and Simon King, *Impact of different speech types on listening effort*, Interspeech 2018, Hyderabad, India

Wavenet based SSDRC: wSSDRC, (S. Muhammed et al. 2018)



- Similar to Rethage et al.: A Wavenet for speech denoising, ICASSP2018

Sound Examples

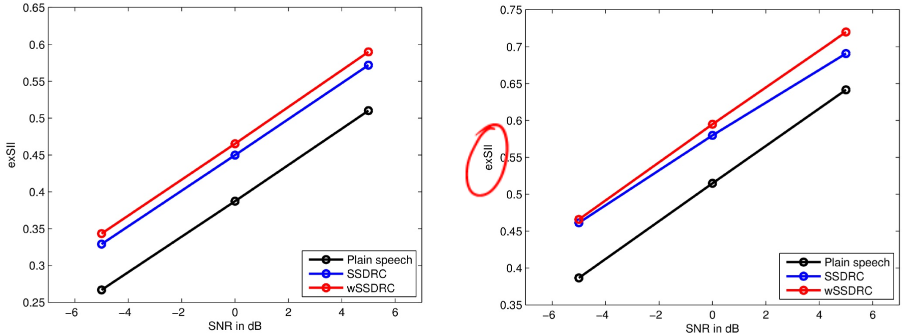


clean speech

SSDRC

	1	2	3
Plain			
SSDRC			
wSSDRC			

Objective evaluations



- **Left:** with stationary white noise (SWN); • **Right:** with stationary shaped noise (SSN)

Listening effort (O. Simantiraki et al. 2018)

- Listening Effort: "The mental exertion required to attend to and understand, an auditory message." *McGarrigle et al*
- Self-reports
- Behavioural measures (single/dual-task → reaction time)
- Physiological measures (fMRI, EEG, skin conductance, heart rate, muscle tension, pupil size, hormone levels)

Listening effort (O. Simantiraki et al. 2018)

- Listening Effort: "The mental exertion required to attend to and understand, an auditory message." *McGarrigle et al*
- Self-reports
- Behavioural measures (single/dual-task $\rightarrow$ reaction time)
- Physiological measures (fMRI, EEG, skin conductance, heart rate, muscle tension, pupil size, hormone levels)

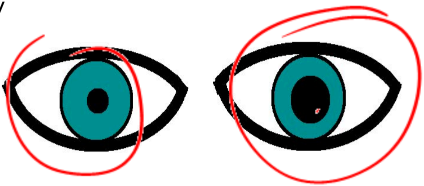
Listening effort (O. Simantiraki et al. 2018)

- Listening Effort: "The mental exertion required to attend to and understand, an auditory message." *McGarrigle et al*

- Self-reports
- Behavioural measures (single/dual-task $\rightarrow$ reaction time)
- Physiological measures (fMRI, EEG, skin conductance, heart rate, muscle tension, pupil size, hormone levels)

Pupillometry

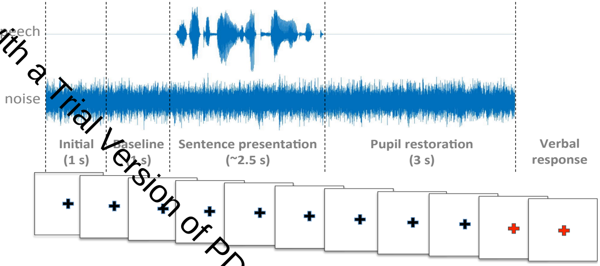
- Pupil Dilation:
  - Widely used as a measure of mental effort
  - More challenging listening conditions $\rightarrow$ Larger pupil size
  - Sensitive to differences in speech intelligibility, masker type, sentence complexity, location uncertainty, motivation
- Pupil Data:
  - Mean dilation
  - Peak dilation
  - Peak latency



Task experiment

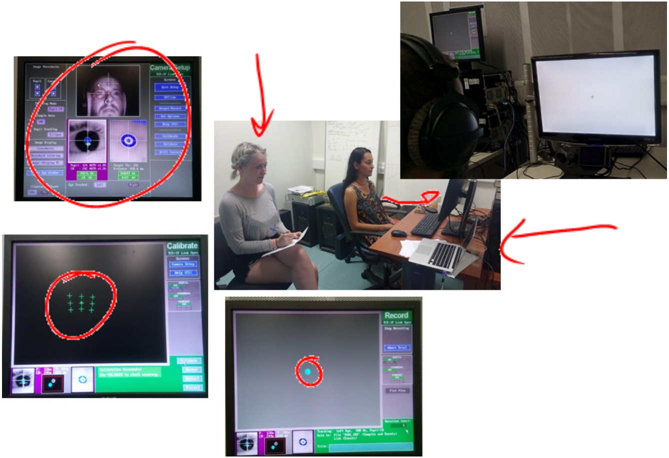
- **Question:** Does listening effort differ among different speech types? Plain, Lombard, Modified speech (SSDRC), Synthetic speech (TTS)
- **Listeners and Design:**
  - 26 young adults (age range 18-24), normal hearing, native British English (3 participants excluded)
  - Harvard sentences - male English talker
  - Speech Shaped Noise at -1, -3 and -5 dB SNR
  - 12 blocks, 20 sentences (first 5 used for familiarisation)
  - Audiological screening (hearing test)
  - Whole procedure with 5-min break took approximately 1h

Experiment setup



- **Task**: Try to recall as many words as you can
- **Data Collected**
  - **Pupil size** (EyeLink 1000)
  - **Intelligibility scores** (% correct words)
  - **Subjective rating**: *"How much effort did it take to listen and understand the sentences in this block?"*. Continuous scale from 0 to 10

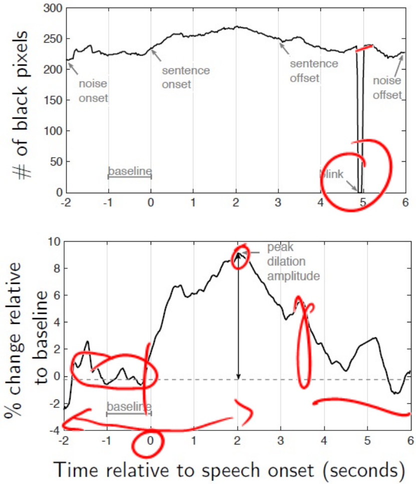
Experiment setup



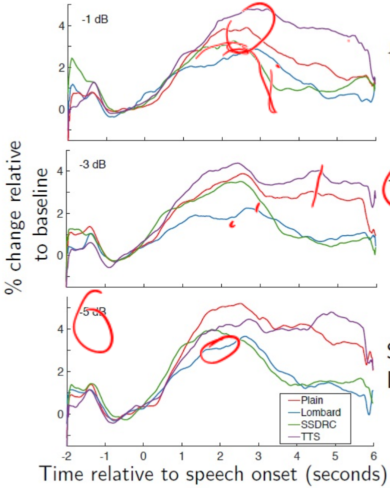
Pupil Data

- **Preprocessing:**
  - 5 first traces of each block were excluded
  - Downsampling to 50 Hz
  - Pupil size measured in units of area was converted to diameter
  - Blink detection and computation of the percentage of blinks (traces were excluded when blinks were more than 15%)
  - Linear interpolation from the start to the end of the blink
  - 5-point moving average smoothing filter
  - Pupil data calibration (proportional increase in pupil dilation relative to the baseline)
  - Average of the traces of each block

Example of pre-processing



some RESULTS



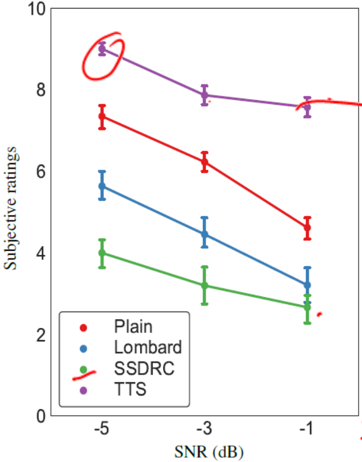
{Lombard,SSDRC,Plain} - TTS *

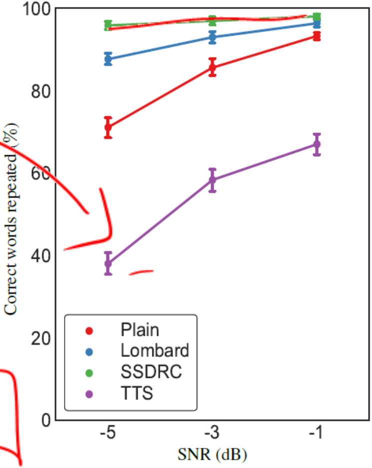
{Lombard,SSDRC} - {Plain,TTS} *

SSDRC - {Plain,TTS} *

Lombard - Plain *

Subjective effort & intelligibility





Outline

R. J. Niederjohn and J. H. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 24, pp. 277–282, Aug. 1976.

B. Sauert and P. Vary, "Near end listeniing enhancement: Speech intelligibility improvement in noisy environments," in *Proceedings of IEEE ICASSP-2006*, (Toulouse, France), pp. 493–496.

B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *Proc. of ITG-Fachtagung Sprachkommunikation*, vol. 9, (Berlin [u.a.]), VDE-Verlag, 10 2010.

B. Sauert and P. Vary, "Near end listening enhancement considering thermal limit of mobile phone loudspeakers," in *Proc. of Conf. on Elektronische Sprachsignalverarbeitung (ESSV)*, vol. 61, (Dresden, Germany), pp. 333–340, ITG, DEGA, TuDPress Verlag der Wissenschaften GmbH, 9 2011.

Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Interspeech 2012*, 2012.

C. H. Taal, R. C. Hendriks, and R. Heusdens, "A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure," in *Proceedings of IEEE-ICASSP*, pp. 4061–4064, Mar. 2012.

V. Hazan and A. Simpson, "Cue-enhancement strategies for natural VCV and sentence materials presented in noise," *Speech, Hearing and Language*, vol. 9, pp. 43–55, 1996.

S. D. Yoo, J. R. Boston, A. El-Jaroudi, C.-C. Li, J. D. Durrant, K. Kovacyk, and S. Shaiman, "Speech signal modification to increase intelligibility in noisy environments," *J. Acoust. Soc. Am.*, vol. 122, pp. 1138–1149, Aug. 2007.

B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *ITG-Fachtagung Sprachkommunikation*, (Bochum, Germany), 2010.

T. C. Zorilă, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Interspeech 2012*, (Portland, USA), 2012.

C. Valentini-Botinhao, J. Yamagishi, S. King, and Y. Stylianou, “Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of hmmbased synthetic speech in noise,” in *Proc. Interspeech*, 2013.

R. Takou, N. Seiyama, and A. Imai, “Improvement of speech intelligibility by reallocation of spectral energy,” in *Proc. Interspeech*, pp. 3605–3607, 2013.

L. Turicchia and R. Sarpeshkar, “A bio-inspired companding strategy for spectral enhancement,” *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 2, pp. 243–253, 2005.

T. Zorilă and Y. Stylianou, “A fast algorithm for improved intelligibility of speech-in-noise based on frequency and time domain energy reallocation,” in *Proc. Interspeech*, pp. 60–64, 2015.

P. Petkov and W. Kleijn, “Spectral dynamics recovery for enhanced speech intelligibility in noise,” *IEEE Trans. Audio Speech Language Process.*, vol. 23, no. 2, pp. 327–338, 2015.

E. Godoy, M. Koutsogiannaki, and Y. Stylianou, “Approaching speech intelligibility enhancement with inspiration from Lombard and clear speaking styles,” *Computer Speech & Language*, vol. 28, no. 2, pp. 629–647, 2014.

D. Erro, C. Zorilă, and Y. Stylianou, “Enhancing the intelligibility of statistically generated synthetic speech by means of noise-independent modifications,” *IEEE Trans. Audio Speech Language Process.*, vol. 22, no. 12, pp. 2101–2111, 2014.

S. Flanagan, T. Zorilă, Y. Stylianou, and B. Moore, “Speech processing to improve the perception of speech in background noise for children with auditory processing disorder and typically developing peers,” *Trends in Hearing*, vol. 22, pp. 1–8, 2018.

A. Griffin, T. Zorilă, and Y. Stylianou, “Improved face-to-face communication using noise reduction and speech intelligibility enhancement,” in *Proc. IEEE ICASSP*, pp. 5103–5107, 2015.

V. Tsiaras, C. Zorilă, Y. Stylianou, and M. Akamine, “Real time speech-in-noise intelligibility enhancement based on spectral shaping and dynamic range compression,” in *ICASSP Show & Tell Session*, 2014.

THANK YOU

for your attention