“What's Happening in Speech Enhancement and Acoustic Signal Processing?”

UK-Speech, September 2013

Khan Baykaner, University of Surrey
Andrew Hines, Trinity College Dublin
Alastair Moore, Imperial College London

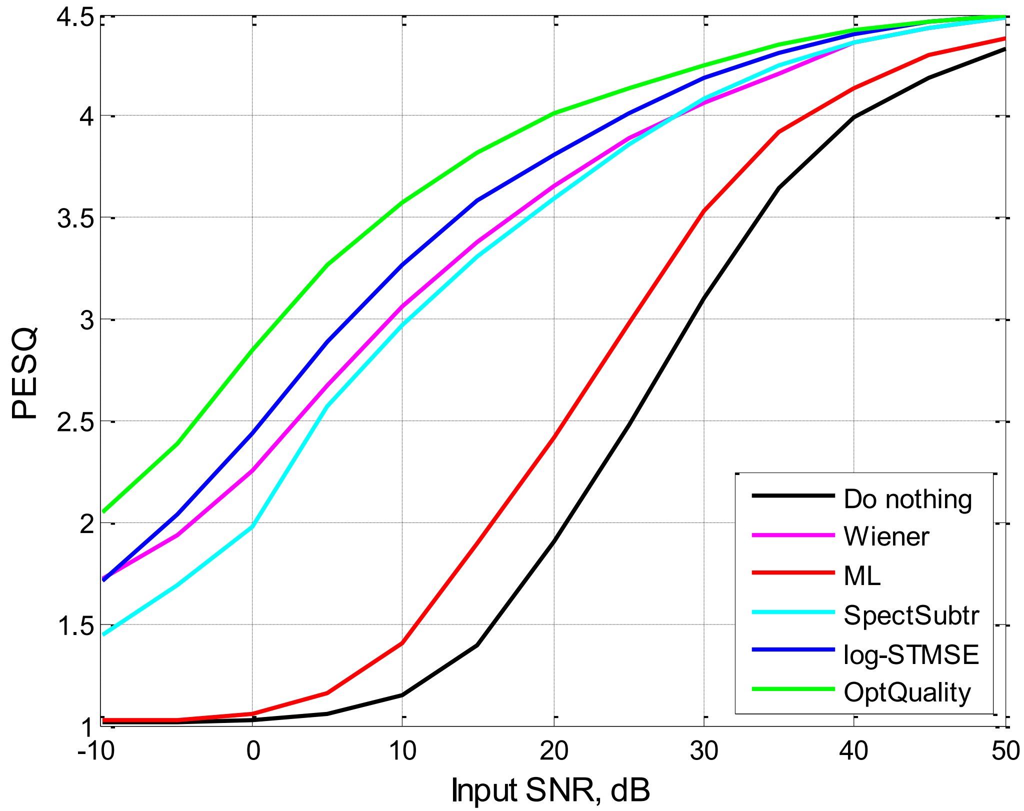
Convener: Patrick A. Naylor p.naylor@imperial.ac.uk

Overview

- Noise reduction
- Speech Enhancement using multichannel speech input and microphone arrays
- Dereverberation
- Instrumental speech quality estimation
- Effect of noise reduction on intelligibility

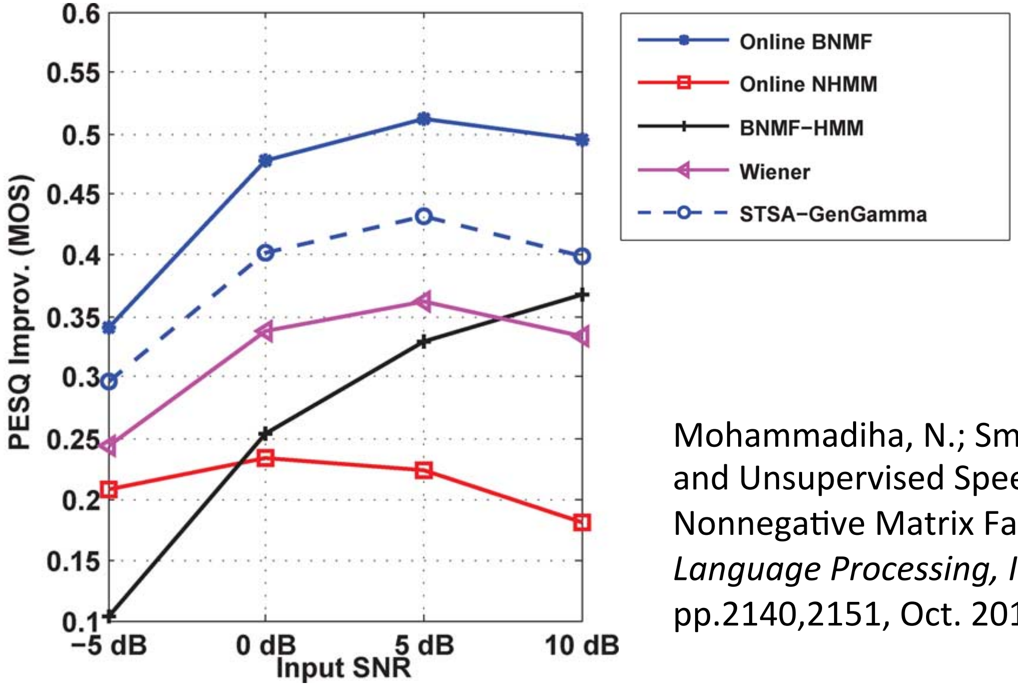
Noise Reduction

- Machine learning
  - Classic approaches apply a time-varying filter (freq. domain gain modification), designed using rules employing Gaussian or super-Gaussian models.
  - Machine learning approaches aim to learn the rule from training data
    - Measure the a priori and a posteriori SNR and deduce the gain rule relating them
    - Shows PESQ improvements of 0.1 to 0.2 compared to logMMSE



I. Tashev and M. Slaney, "Data Driven Suppression Rule for Speech Enhancement", in Proc. Information Theory and Applications Workshop, UCSD, Feb 2013.

- Model-based speech enhancement / NMF
  - Supervised algorithms based on HMMs can work well but need an a priori model for each noise type
  - New methods exploit nonnegative matrix factorization (NMF) in both supervised and unsupervised forms



Mohammadiha, N.; Smaragdis, P.; Leijon, A., "Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.21, no.10, pp.2140,2151, Oct. 2013

Multichannel Speech Input

- Hardware examples showing some illustration of configurations

- AMI

- Eigenmike

- Meeting transcription (NTT)

- Smartphone

- Dataset examples
  - AMI Corpus: Meeting corpus, simultaneous array and close mic recordings
  - CMU Robust Speech Recognition Group: Microphone Array Database
  - Multi-channel Overlapping Numbers Corpus (Idiap)
  - Reverb Challenge datasets
    - http://reverb2014.dereverberation.com/data.html

- Room Impulse Responses
  - AcouSP
    - Portal to several databases of room impulse response measurements
    - www.commsp.ee.ic.ac.uk/~acousp

http://reverb2014.dereverberation.com

REVERB CHALLENGE

Home | Introduction | Data | Enhancement task | ASR task | Instructions | Download | Workshop

Welcome to the REVERB challenge

Recently, substantial progress has been made in the field of reverberant speech signal processing, including both single- and multi-channel de-reverberation techniques, and automatic speech recognition (ASR) techniques robust to reverberation. To evaluate state-of-the-art algorithms and draw new insights regarding potential future research directions, we are now launching and calling for participation* in the **REVERB (REverberant Voice Enhancement and Recognition Benchmark) challenge** that provides an opportunity to the researchers in the field to carry out a comprehensive evaluation of their methods based on a common database and on common evaluation metrics. This is a multidisciplinary challenge. We encourage participants from both the speech enhancement and speech recognition communities. All entrants will be invited to submit papers describing their work to a dedicated **workshop held in conjunction with ICASSP 2014 and HSCMA 2014**.

*PDF version of call for participation is available here.

Important dates

Jul 1, 2013
Release of development dataset and scripts for evaluation

Nov 5, 2013
Release of evaluation dataset

Dec 1, 2013
Deadline for submission of results

Jan 10, 2014
Deadline for submission of papers

Feb 28, 2014
Notification of acceptance

May 10, 2014
Workshop in conjunction with

Microphone Array Processing

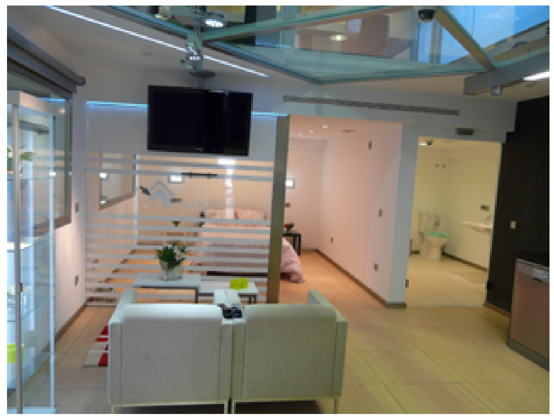
- "The adaptation of beamforming methods to speech enhancement problems remains an open issue. These difficulties may be attributed to the wide-band and nonstationary characteristics of a speech signal and to the very long, typically time-varying, room impulse responses (RIRs) relating moving speakers and microphones in acoustic enclosures."
  - Sharon Gannot

Existing Approaches

- Fixed beamforming
  - Combine the microphone signals using a time-invariant filter-and-sum operation (data-independent)
    - [Jan and Flanagan, 1996]; [Doclo and Moonen, 2003].
- Blind Source Separation (BSS)
  - Considers the received signals at the microphones as a mixture of all sound sources filtered by the RIRs. Utilizes Independent Component Analysis (ICA) techniques
    - [Makino et al., 2007]; TRINICON, [Buchner et al., 2004].
- Adaptive Beamforming
  - Combine the spatial focusing of fixed beamformers with adaptive suppression of (spectrally and spatially time-varying) background noise
    - [Cox et al., 1987]; [Van Veen and Buckley, 1988]; [Van Trees, 2002].
- Computational Auditory Scene Analysis (CASA)
  - Aims at performing sound segregation by modelling the human auditory perceptual processing
    - [Wang and Brown, 2006].

Ad hoc arrays using wireless acoustic sensor networks

- Advantages of ad hoc wireless microphone arrays (WASN)
  - No calibration needed
  - Better sampling of more of the sound field, given enough mics
  - Easy deployment
- Applications
  - Cooperative hearing aids
  - Smart homes
  - Surveillance



- References

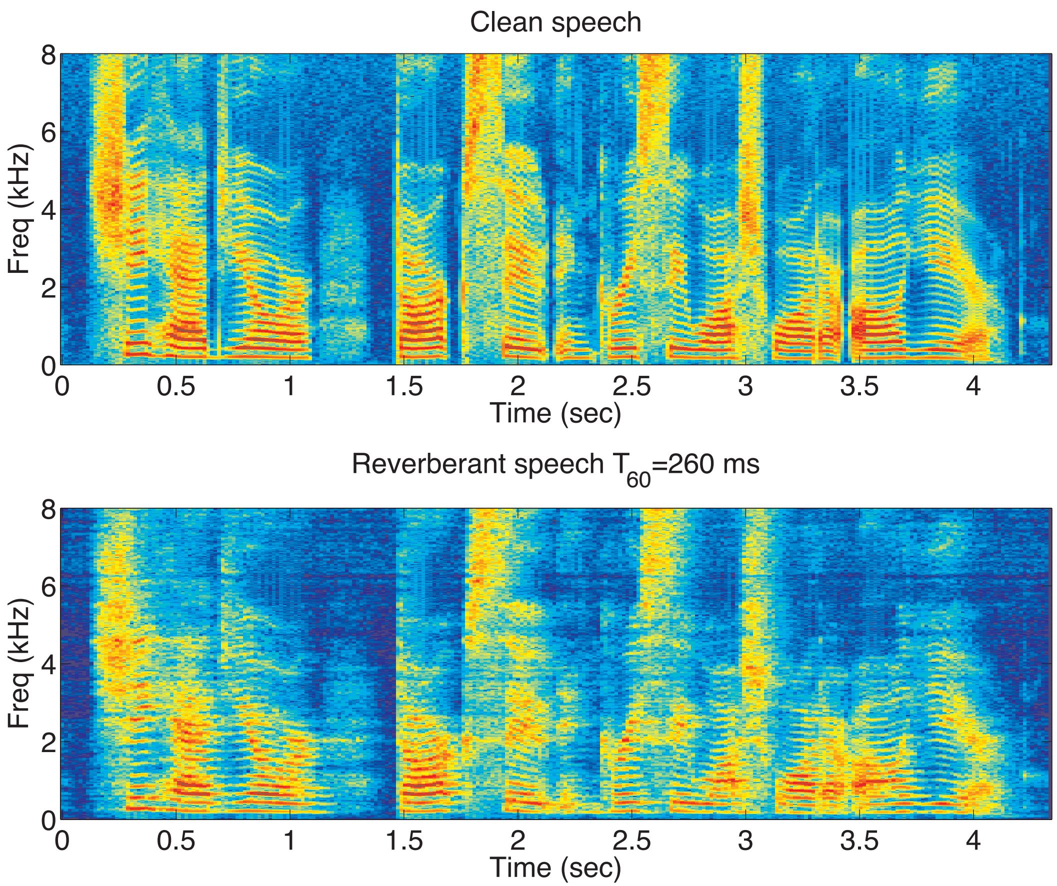
A. Bertrand and M. Moonen, “Distributed LCMV beamforming in a wireless sensor network with single-channel per-node signal transmission,” IEEE Transactions on Signal Processing, 61:3447–3459, 2013

S. Markovich-Golan, S. Gannot, and I. Cohen, ”Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks”, IEEE Transactions on Audio, Speech, and Language Processing, 21(2):343–356, 2013.

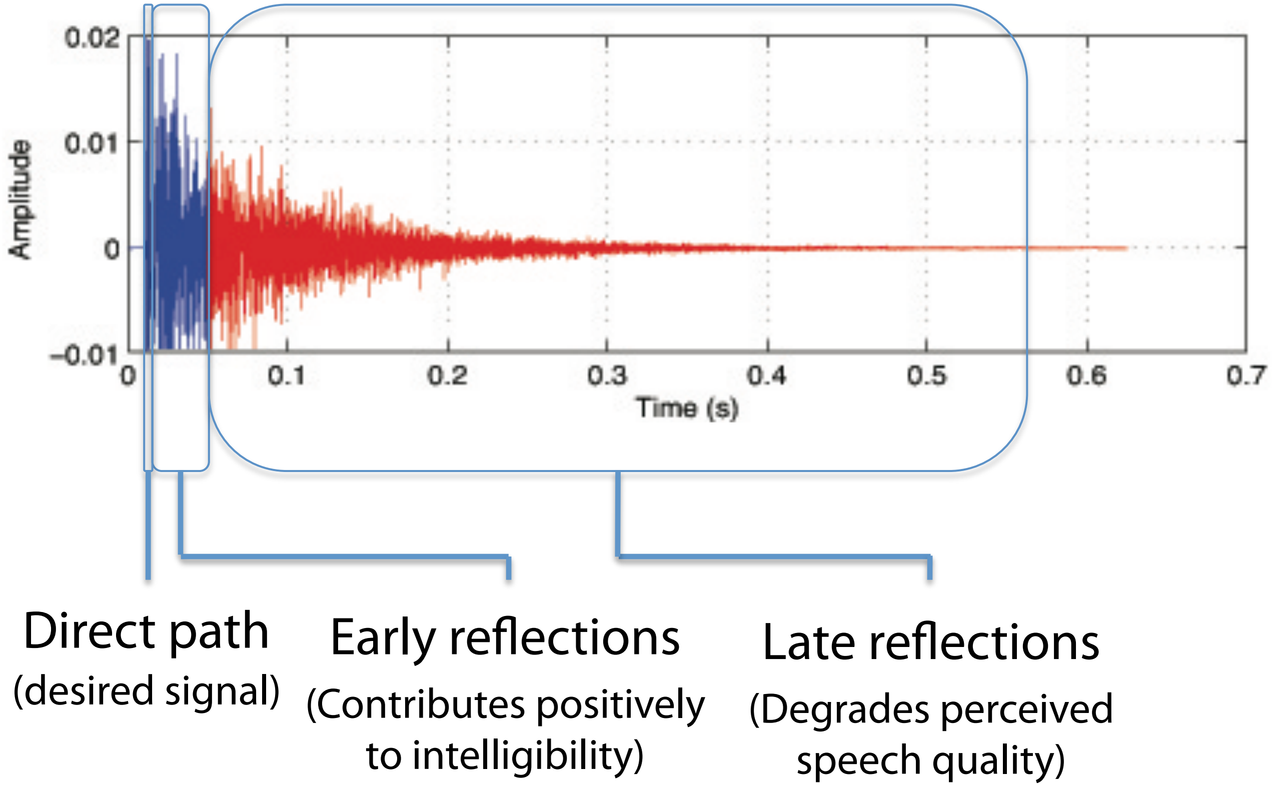
J. Szurley, A. Bertrand, and M. Moonen, “Improved tracking performance for distributed node-specific signal enhancement in wireless acoustic sensor networks”, in Proc. ICASSP 2013.

Dereverberation

- Reverberation is the convolution of the room impulse response (RIR) with the desired speech signal

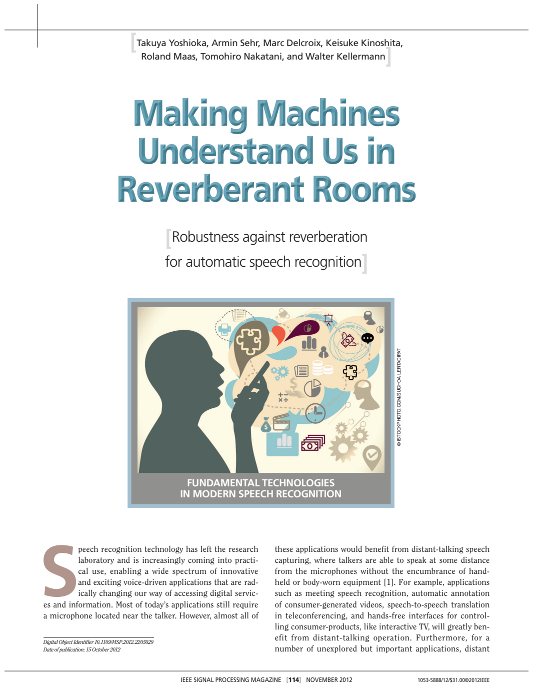


Dereverberation



- Aim of dereverberation is to remove at least the reverberation tail and possibly also early reflections

Reviews



Takuya Yoshioka, Armin Sehr, Marc Delcroix, Keisuke Kinoshita,
Roland Maas, Tomohiro Nakatani, and Walter Kellermann

Making Machines Understand Us in Reverberant Rooms

Robustness against reverberation for automatic speech recognition

FUNDAMENTAL TECHNOLOGIES IN MODERN SPEECH RECOGNITION

Speech recognition technology has left the research laboratory and is increasingly coming into practical use, enabling a wide spectrum of innovative and exciting voice-driven applications that are radically changing our way of accessing digital services and information. Most of today's applications still require a microphone located near the talker. However, almost all of these applications would benefit from distant-talking speech capturing, where talkers are able to speak at some distance from the microphones without the encumbrance of hand-held or body-worn equipment [1]. For example, applications such as meeting speech recognition, automatic annotation of consumer-generated videos, speech-to-speech translation in teleconferencing, and hands-free interfaces for controlling consumer-products, like interactive TV, will greatly benefit from distant-talking operation. Furthermore, for a number of unexplored but important applications, distant

Digital Object Identifier 10.1109/MSP.2012.2205029
Date of publication: 15 October 2012

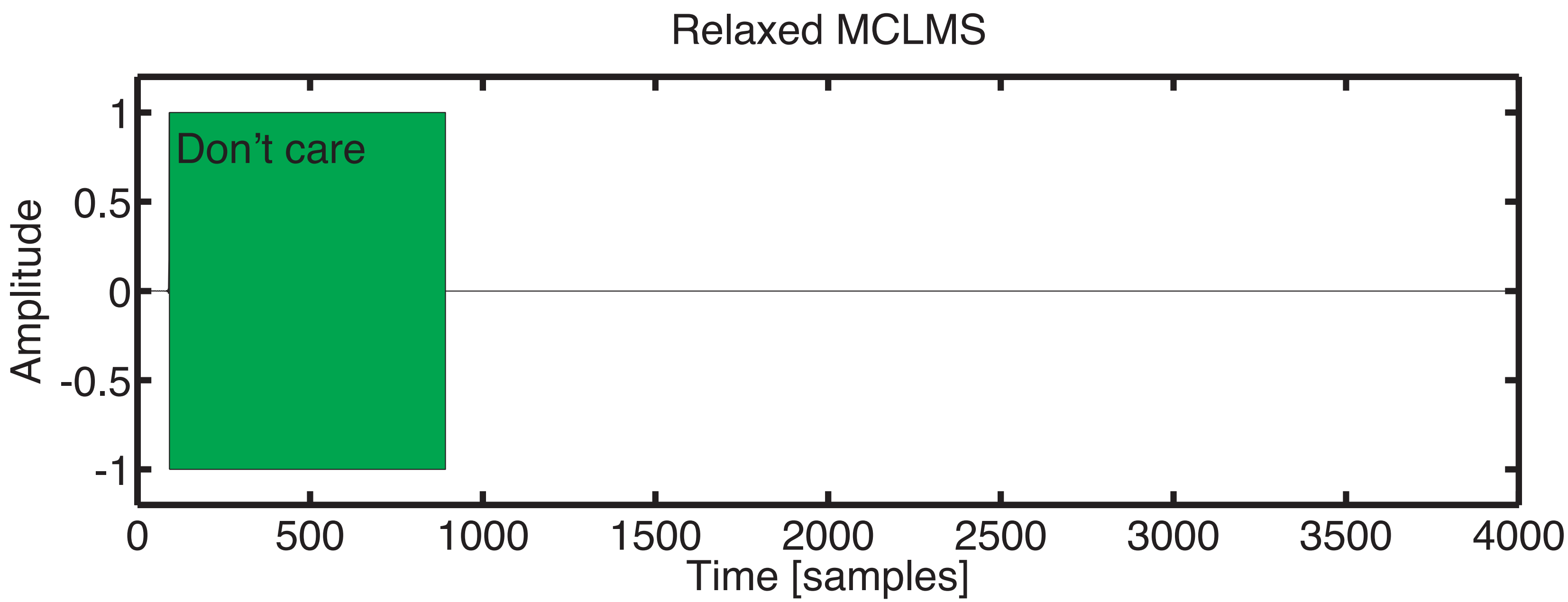
IEEE SIGNAL PROCESSING MAGAZINE [114] NOVEMBER 2012 1053-5888/12/$31.00©2012IEEE

[Yoshioka et al, 2012]

[Naylor and Gaubitch, 2010]

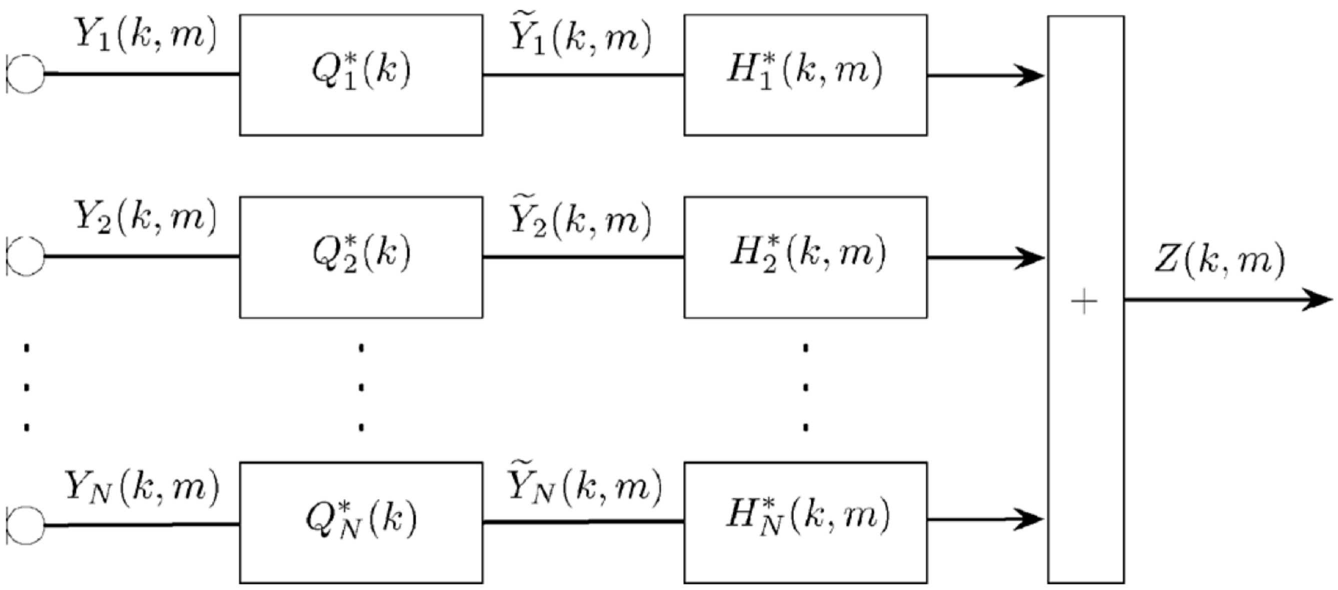
Channel equalisation

- Aim: Design a linear filter to equalise magnitude and phase
- Current research is looking at how to define the target equalised response to maximise robustness to system identification errors whilst maintaining quality [Lim and Naylor, 2013], [Kodrasi and Doclo, 2013]



Beamforming I

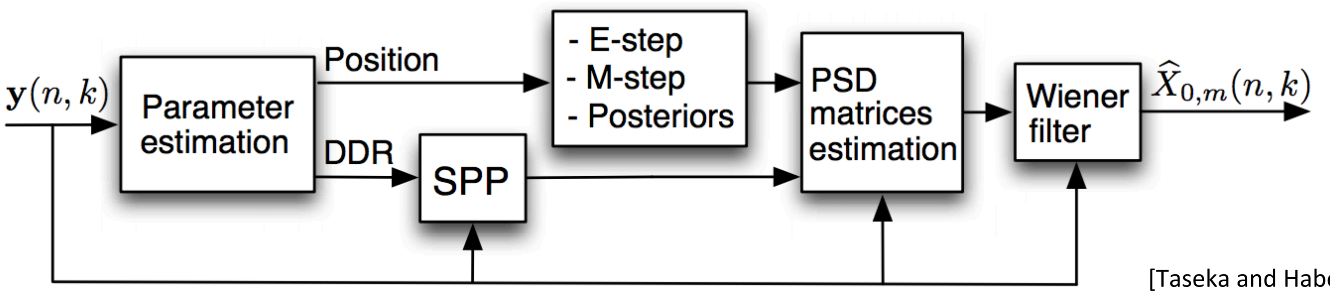
- Aim: Select the signal coming from a particular direction
- Requires multiple microphones
- Spatial filter uses signals from all channels to extract the desired signal
- Remove residual decay and noise using spectral enhancement [Habets and Benesty, 2013]



[Habets and Benesty, 2013]

Beamforming II

- Time varying spatial filter uses estimates of the direction of arrival and power spectral density of the desired source(s) and incorporates an arbitrary spatial response [Thiergart et al, 2013]
- For moving sources, online direction of arrival estimates using expectation maximisation looks promising, at least for modest amounts of reverbaration [Taseka and Habets, 2013]



[Taseka and Habets, 2013]

Bibliography

Dereverberation

[Habets and Benesty, 2013] A two-stage beamforming approach for noise reduction and dereverberation, IEEE Audio, Speech, Language Process., vol. 21, no. 5, pp. 945-958, 2013.

[Kodrasi and Doclo, 2013] Regularized subspace-based acoustic multichannel equalization for speech dereverberation, Proc. European Signal Processing Conference (EUSIPCO), Sep. 2013

[Lim and Naylor, 2013] Robust speech dereverberation using subband multichannel least squares with variable relaxation, Proc. European Signal Processing Conference (EUSIPCO), Sep. 2013

[Naylor and Gaubitch, 2010] Speech dereverberation. Springer, 2010.

[Taseka and Habets, 2013] An online EM algorithm for source extraction using distributed microphone arrays, Proc. European Signal Processing Conference (EUSIPCO), Sep. 2013

[Thiergart et al, 2013] An informed MMSE filter based on multiple instantaneous direction-of-arrival estimates, Proc. European Signal Processing Conference (EUSIPCO), Sep. 2013

[Yoshioka et al, 2012] Making Machines Understand Us in Reverberant Rooms: Robustness Against Reverberation for Automatic Speech Recognition, Signal Processing Magazine, IEEE , vol.29, no.6, pp.114,126, Nov. 2012

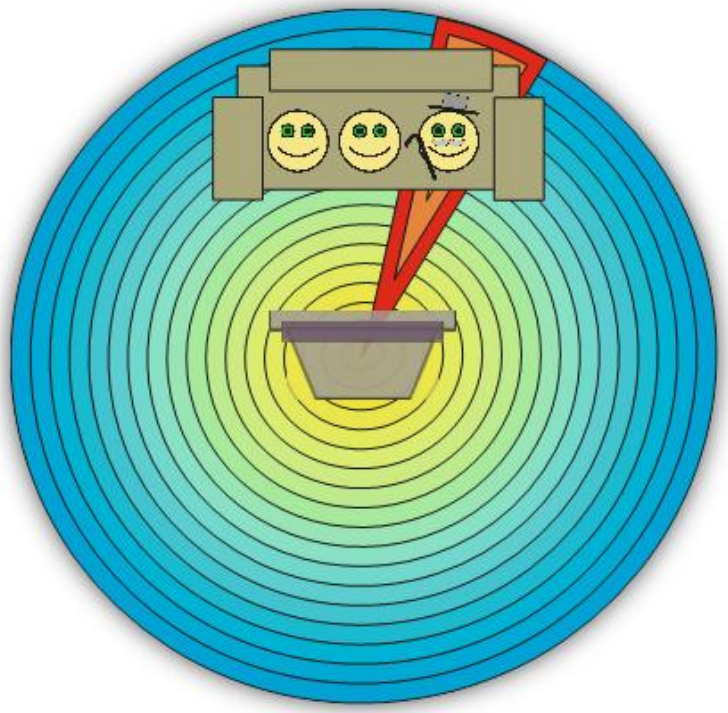
Intelligibility Prediction

Recent Methods	Description	Application
Hearing Aid Speech Quality Index–Intelligibility (HASQI-I) (Kates, 2013; Kates & Arehart, 2010)	Auditory Model + correlation (intrusive) – aims to keep computational costs low / implement in hardware	Generic / Hearing aids
NSIM (Hines et al., 2010)	Auditory Model + similarity metric (intrusive)	Generic / Hearing aids
(Christiansen et al. 2010)	Auditory Model + correlation (intrusive)	Generic / time-frequency weighted noise
Fractional AI (Louizou & Ma, 2011)	Modification to the articulation index to allow for prediction of non-linearly amplified audio	time-frequency weighted noise (NR)
STOI (Taal et al., 2011)	Simplified auditory model + correlation	Real-time / Generic / time-frequency weighted noise (NR)
Multi-sEPSM (Jogensen et al., 2013)	Auditory Model with focus on envelope SNR	Generic / non-stationary interferers

- Strong focus on current models to interpret the output of a model of the auditory periphery
- Also a strong focus on current models to predict intelligibility of audio programmes featuring both linear & non-linear processing (caused by noise reduction algorithms).
- In part, this is driven by the application of noise reduction algorithms to hearing aids.

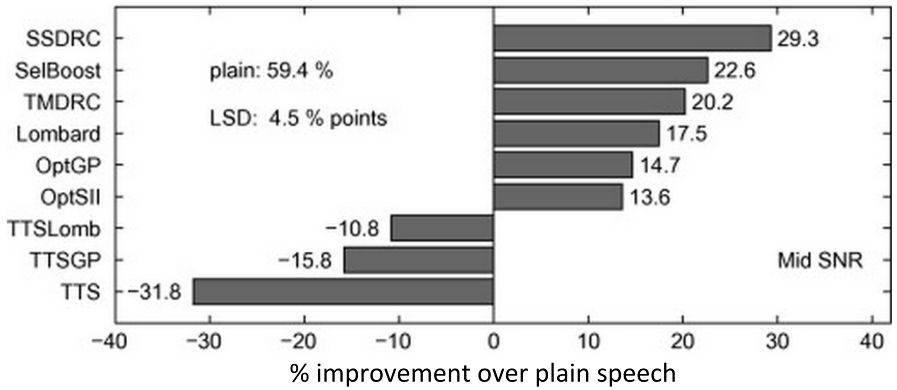
Intelligibility improvement via Personal Audio

- Using a superdirective array to strengthen high frequencies over a small region, intelligibility can be improved for the hearing impaired while not affecting normal hearing listeners (Galvez & Elliot 2013)
- 10-15 dB contrast for 1-8kHz using 4x8 array of hypercardioid loudspeakers.

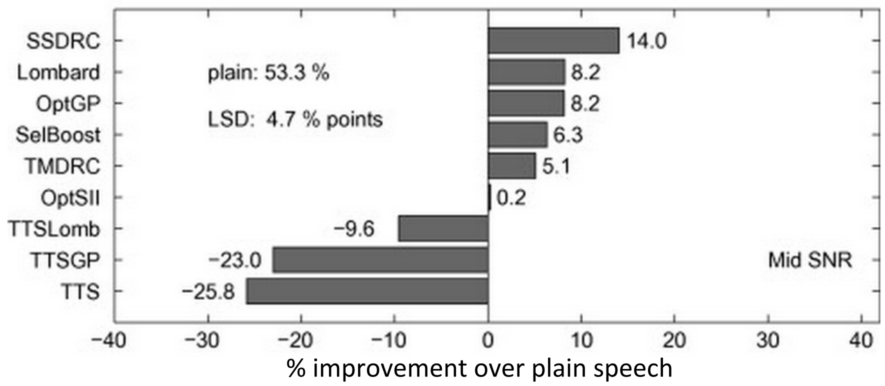


Modifying Speech to Boost Intelligibility

- Apply a noise shaped (in frequency) gain function (Sauert & Vary. 2006)
- Modifications in time and frequency by:
  - Using a harmonic speech model & dynamic range compression (Erro et al., 2012)
  - Optimising for a perceptual distortion metric based on an auditory model (Taal et al., 2012)
  - Spectral shaping and dynamic range compression (Zorila et al., 2012)
- Comparison of modification methods showed that speech pre-processing can enhance intelligibility more effectively than Lombard speech (Cooke et al. 2013)



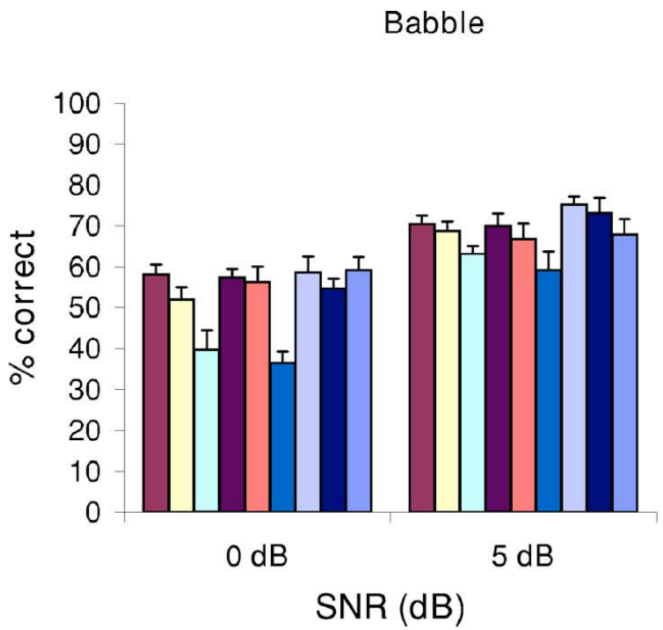
Speech Shaped Noise

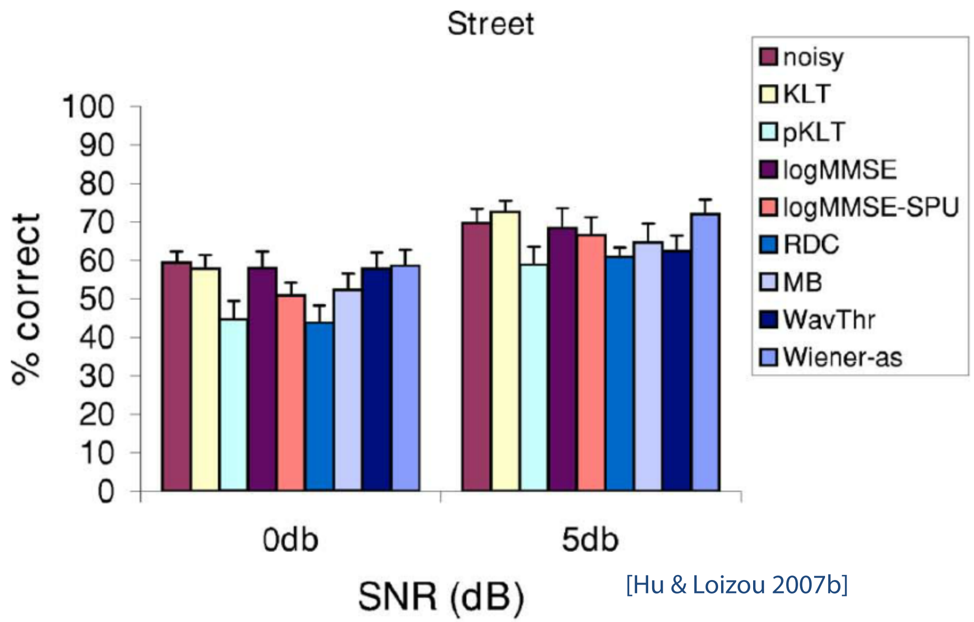


Competing Speech

Speech Intelligibility – Noise reduction

- Some noise reduction algorithms are deleterious to intelligibility (Hu & Loizou, 2007a&b Li et al., 2011)
- "one reason that existing algorithms do not improve speech intelligibility is because they allow amplification distortions in excess of 6 dB" (Kim & Loizou, 2011)
- Spectral Subtraction and Minimum Mean Squared Error Spectral Subtraction reduced intelligibility, and Subspace Enhancement had no effect (Hilkhuysen et al., 2012)
- (ideal) Time-Frequency masking improves intelligibility but at the cost of quality (e.g. by introducing musical noise) (Brons et al., 2012)





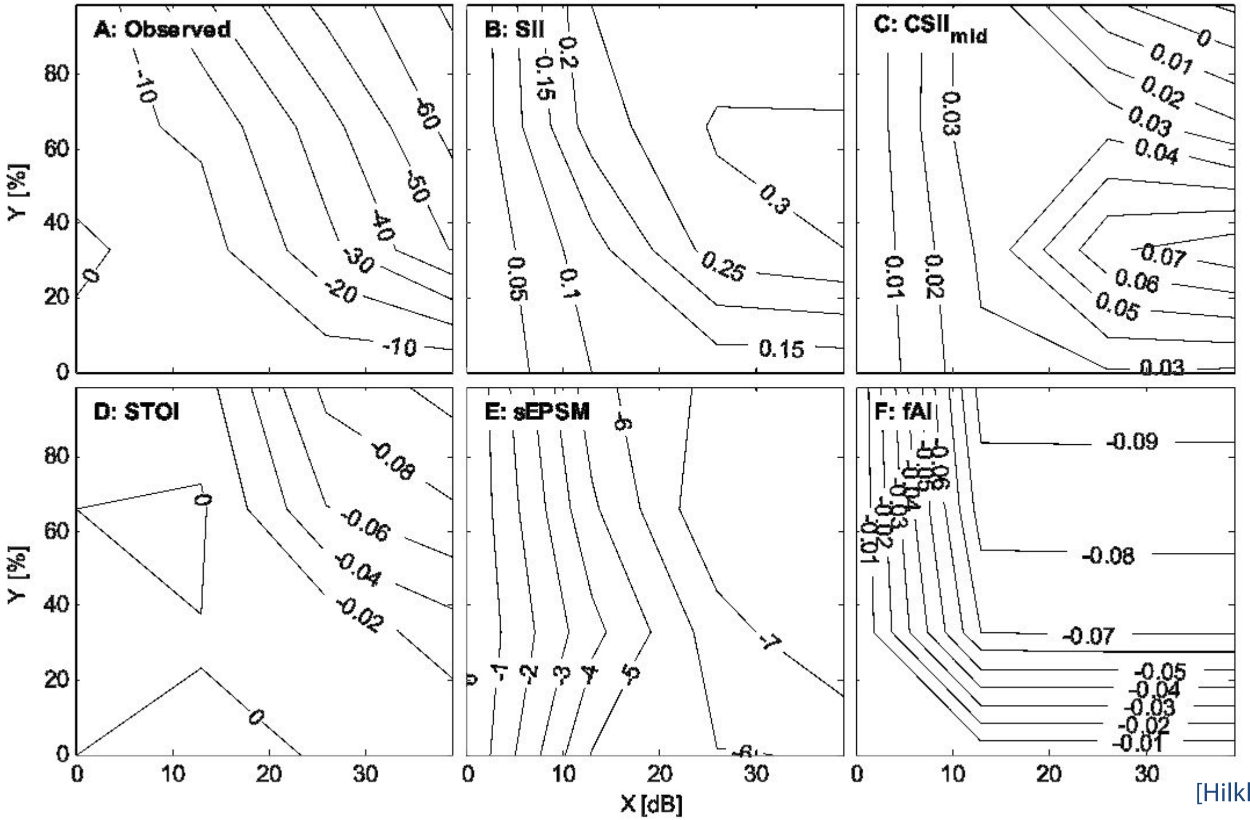
[Hu & Loizou 2007b]

Speech Intelligibility – Noise reduction

- Use of intelligibility models in the design stage of noise reduction algorithms. Five intelligibility prediction models tested in Hilkhuysen & Huckvale (2013)

SII (ANSI 1997) **CSII** (Kates & Arehart, 2005) **STOI** (Taal et al., 2011) **sEPSM** (Jorgensen & Dau, 2011) **fAI** (Loizou & Ma, 2011)

- Predictions compared with subjective listening test - only fAI identified the optimal NR parameters (and not uniquely)



[Hilkhuysen & Huckvale, 2013]

Intelligibility References

In order of appearance:

Kates, J., 2013; "An auditory model for intelligibility and quality predictions" Proceedings of Meetings on Acoustics vol. 19

Kates, J., and Arehart, K., 2010; "The hearing aid speech quality index (HASQI)" Journal of the Audio Engineering Society vol. 58

Christiansen, C., and Pedersen, M. S., and Dau, T., 2010: "Prediction of speech intelligibility based on an auditory preprocessing model" Speech Communication vol. 52

Hines, A., and Harte, N.; 2012; "Speech intelligibility prediction using a Neurogram Similarity Index Measure" Speech Communication, 54 (2)

Loizou, P. C. and Ma, J. (2011). "Extending the Articulation Index to Account for Non-Linear Distortions Introduced by Noise-Suppression Algorithms," J. Acoust. Soc. Am. 130, 986-995.

Taal, C., and Hendriks, R., and Heusdens, R., and Jesper, J., 2011; "An alogirthm for intelligibility prediction of time-frequency weighted noisy speech" IEEE Transactions on Audio, Speech, and Language Processing 19

Jorgensen, S., and Ewert, S. D., and Dau, T., 2013; "A multi-resolution envelope-power based model for speech intelligibility", Journal of the Acoustical Society of America, vol. 134

Galvez, Marcos F. S. & Elliot, Stephen J. 2013: "The design of a personal audio superdirective array in a room", AES 52nd international conference, September

Sauert, Bastian, and Vary, Peter, 2006; "Near end listening enhancement: Speech intelligibility improvement in noisy environments", International Conference on Acoustics, Speech, and Signal Processing

Erro, Daniel, and Stylianou, Yannis, and Navas, Eva, and Hernaez, Inma, 2012; "Implementation of simple spectral techniques to enhance the intelligibility of speech using a harmonic model", 13th Annual conference of the International Speech communication association

Taal, Cees H., and Hendriks, Richard C., and Heusdens, Richard, 2012; "A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure", International Conference on Acoustics, Speech, and Signal Processing

Zorila, Tudor-Catalin, and Kandia, Varvara, and Stylianou, Yannis, 2012; "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression", 13th Annual conference of the International Speech communication association

Cooke, Martin & Mayo, Catherine & Valentini-Botinhao, Cassia & Stylianou, Yannis & Sauert, Bastian & Tang, Yan 2013: "Evaluating the intelligibility benefit of speech modifications in known noise conditions", Speech Communication 4

Hu, Yi & Louizou, Philipos C., 2007a: "A comparative intelligibility study of speech enhancement algorithms", International Congress on Acoustics Speech and Signal Processing

Hu, Yi & Louizou, Philipos C., 2007b: "A comparative intelligibility study of single-microphone noise reduction algorithms", Journal of the Acoustical Society of America, Volume 122 (3)

Li, Junfeg, and Yang, Lin, and Yan, Yonghong, and Hu, Yi, and Akagi, Masato, and Loizou, Philipos C., 2011; "Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English"

Kim, Gibak, and Loizou, Philipos, 2011; "Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms", Journal of the Acoustical Society of America, vol. 130 (3)

Hilkhuysen, Gaston & Gaubitch, Nikolay, & Brookes, Mike & Huckvale, Mark, 2012: "Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios, Journal of the acoutsical society of america

Brons, I., and Houben, R., and Wouter, A. D., 2012; "Perceptual effects of noise reduction by time-frequency masking of noisy speech", Journal of the Acoustical Society of America vol. 132 (4)

Hilkhuysen, Gaston & Huckvale, Mark, 2013: "Can physical metrics identify noise reduction settings that optimize intelligibility?", Proceedings of Meetings on Acoustics, Volume 19

American National Standards Institute, 1997; "Methods for calculation of the speech intellligibility index"

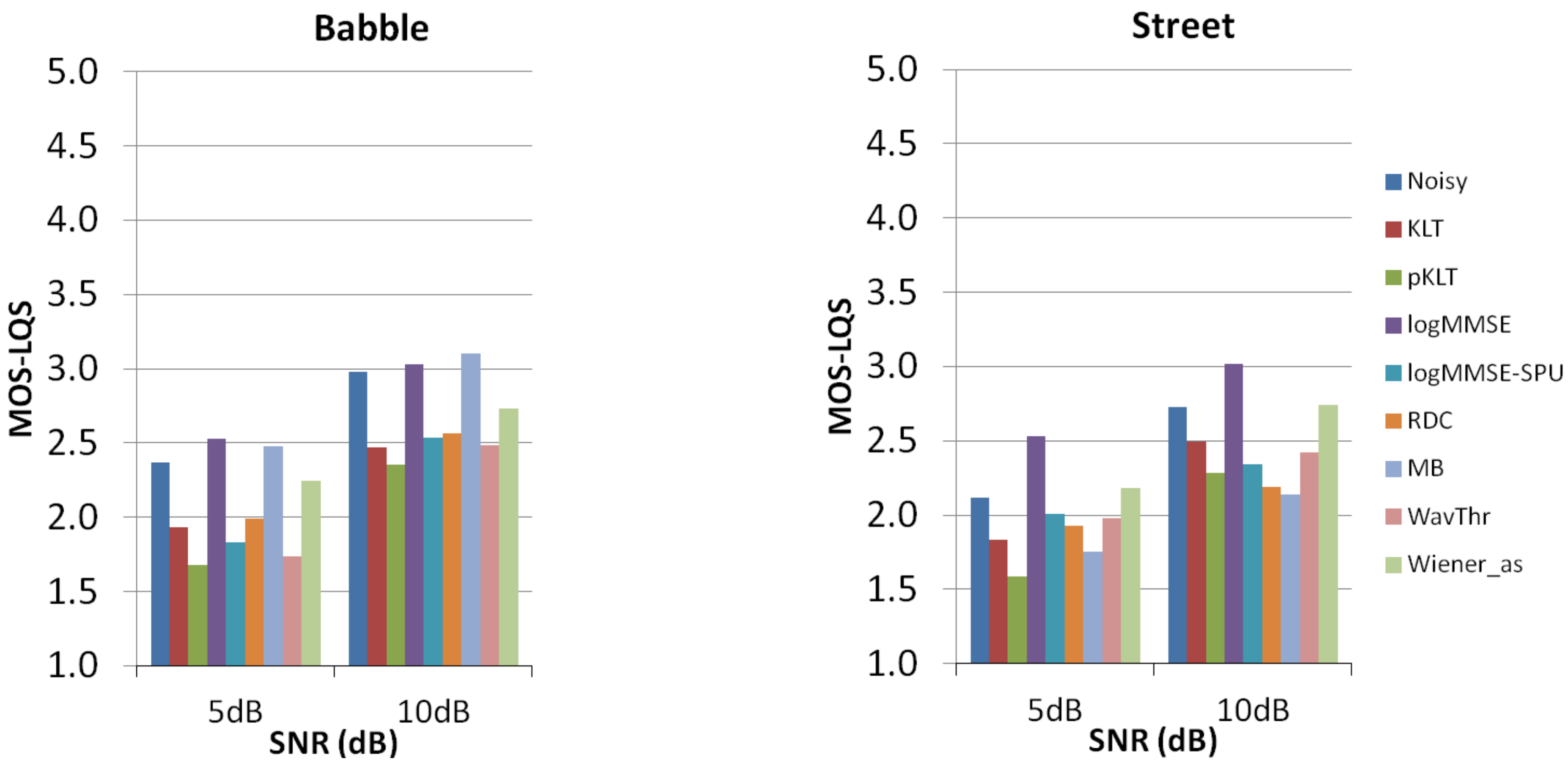
Kates, J., and Arehart, K, 2005; "Coherence and the speech intelligibility index" Journal of the Acoustical Society of America vol. 117 (4)

Jorgensen, S., and Dau, T. (2011), "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing", Journal of the Acoustical Society of America, Vol. 134 (1)

Speech Quality – Noise reduction

- In many scenarios noise reduction algorithms can reduce quality

[Hu & Loizou 2007, Hu & Loizou 2008]



(Adapted from Hu & Loizou, 2007)

Objective Speech Quality Estimation

"Horses for Courses" -
Match the Application to the Model

Application

Plan, optimise, monitor, maintenance

Signal Type

NB/WB/SWB

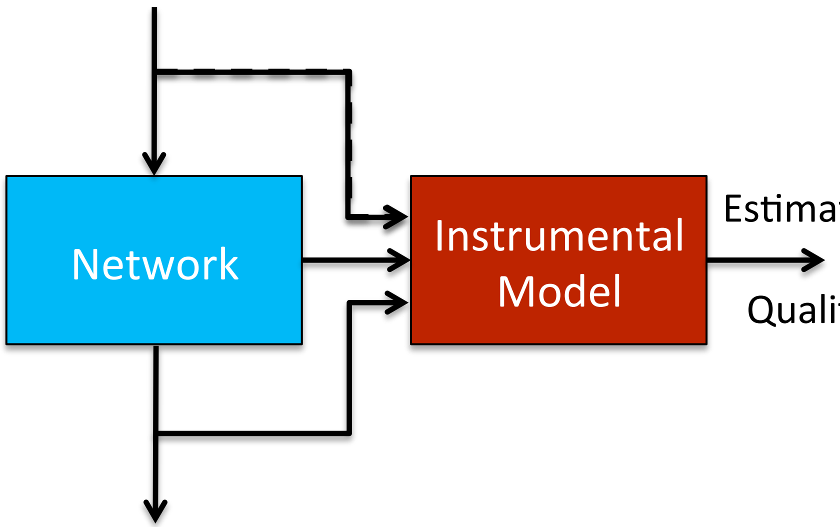
Monaural/binaural

Source of input

Parameter, Simulation, Measurement

Inputs

Params, Ref + Test Sig, Only Test

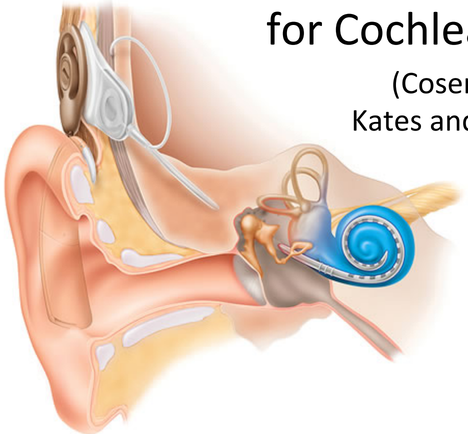


(Figure adapted from Moller et al. 2011)

Examples of Objective Speech Quality Estimation

Speech Quality and Intelligibility for Cochlear Implants

(Cosentino et al.,2013; Kates and Arehart, 2010;)

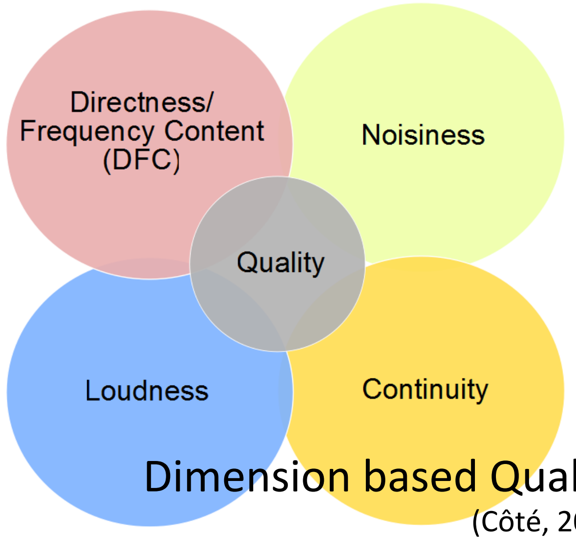


De-reverberated Speech

(Falk, 2010; Naylor, 2010)

Artificial Bandwidth Extension

(Moller et al., 2013)



Dimension based Quality

(Côté, 2011)

Bibliography

Speech Quality Estimation

[ANSI, 2006] "Auditory Non-Intrusive Quality Estimation Plus (Anique+): Perceptual Model for Non-Intrusive Estimation of Narrowband Speech Quality ", ATISPP-0100005.2006, American National Standards Institute, 2006.

[Cosentino et al.,2013] S. Cosentino, T. H. Falk, D. McAlpine, "Predicting the Bilateral Advantage in Cochlear Implantees using a Non-Intrusive Speech Intelligibility Measure," Proc. Interspeech, Lyon, France, Aug. 2013

[Côté, 2011] N. Côté, "Integral and Diagnostic Intrusive Prediction of Speech Quality". Berlin, Springer, 2011.

[Falk, 2010] T. H. Falk, C. Zheng, W.-Y. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," Audio, Speech, and Language Processing, IEEE Transactions on , vol.18, no.7, pp.1766,1774, Sept. 2010

[Grancharov et al., 2006] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," IEEE Audio, Speech, Language Process., vol. 14, no. 6, pp. 1948–1956, 2006.

[Hines et al., 2013] A.Hines, J. Skoglund, A. Kokaram, N.Harte, "Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, 2013.

[Hu and Loizou, 2007] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," Speech Commun., vol. 49, pp. 588–601, 2007.

[Hu and Loizou, 2008] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 16, no. 1, pp. 229–238, 2008.

[ITU-T, 2001] ITU, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Int. Telecomm. Union, Geneva, Switzerland*, ITU-T Rec. P.862 ,* 2001.

[ITU-T, 2004] ITU, "Single-ended method for objective speech quality assessment in narrowband telephony applications," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.563, 2004.

[ITU-T, 2009] ITU, "The E-model: A computational model for use in transmission planning," Int. Telecomm. Uni on, Geneva, Switzerland*, ITU-T Rec. G.107, 2009.*

[ITU-T, 2011] ITU, "Perceptual objective listening quality assessment," Int. Telecomm. Union*, Geneva, Switzerland, ITU-T Rec. P.863, 2011*.

[Kates, 2010] J. Kates, and K. Arehart, "The hearing aid speech quality index (HASQI)" Journal of the Audio Engineering Society vol. 58, 2010

[Moller et al., 2011] S Moller, W-Y Chan, N. Côté, T. H. Falk, A. Raake, and M Waltermann, "Speech quality estimation: Models and trends," Signal Processing Magazine, IEEE, vol. 28, no. 6, pp. 18–28, 2011.

[Moller et al., 2013] S. Moller et al., *Speech Quality Prediction for Artificial Bandwidth Extension Algorithms,* Proc. Interspeech, Lyon, France Aug. 2013

[Naylor and Gaubitch, 2010] P.A. Naylor and N. D. Gaubitch, eds. *"*Speech dereverberation". Springer, 2010.