Enhancing Subjective Speech Intelligibility Using a Statistical Model of Speech

Petko N. Petkov[1], W. Bastiaan Kleijn[1,2], Gustav Eje Henter[1]

[1]Sound and Image Processing Lab, School of Electrical Engineering, KTH – Royal Institute of Technology, Stockholm, Sweden

[2]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand

Overview

Speech in noise can be made more intelligible without an increase in energy!

- Human strategies (Lombard speech) do this to a limited extent;
- Machines can go further but need to be controlled carefully.

An optimization-based framework for speech pre-emphasis may target

- Minimizing a perceptual distortion measure (improving audibility);
- Maximizing recognition accuracy (improving sound discrimination).

We show that the optimization of a measure of speech intelligibility at the text level (based on a clean speech model from ASR) for the parameters of a speech modification strategy can increase intelligibility significantly.

The method requires:

- A phonetic transcription of the message;
- An accurate phone-level waveform segmentation. *A-priori* available in TTS. Extracted for recorded & live speech;
- Disturbance statistics.

Modifications & Intelligibility Optimization

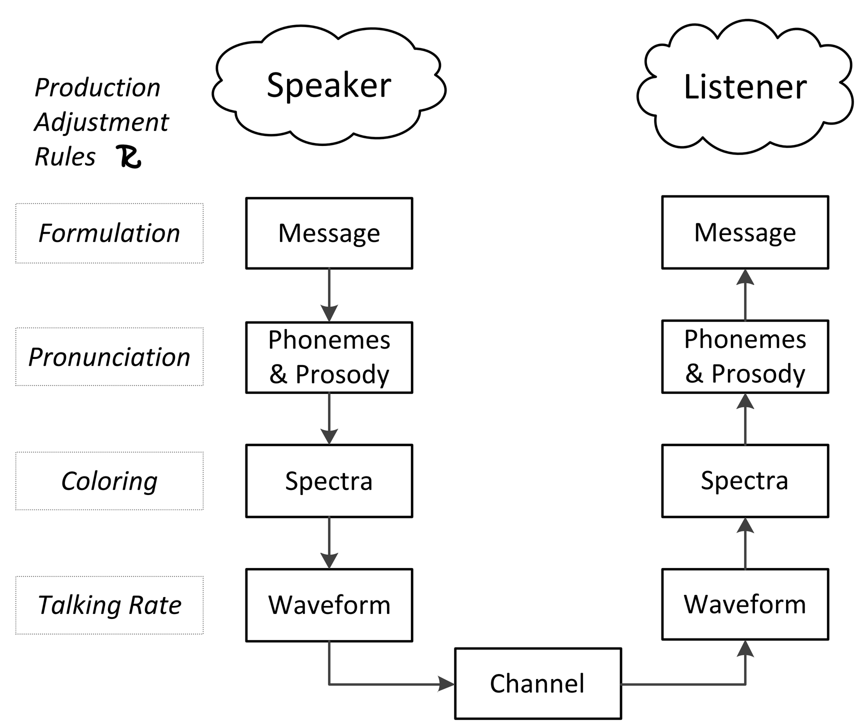

Figure 1: Speech Communication Hierarchy.

Modification-side perspective

- Low-level modifications
  - Efficient but limited;
  - Widely used in practice.
- High-level modifications
  - Need more prior knowledge;
  - Opens wide modificiation space.

Optimization-side perspective

- Low-level measure optimization
  - Low complexity;
  - Modification-specific;
  - Conceptually far from the true objective (message intel.).
- High-level measure optimization
  - More general;
  - Closer to the true objective.

Text-Level Objective Intelligibility

Select the modification parameters c that maximize the probability $p\left(t \mid \widehat{\boldsymbol{F}}_y(c), \mathcal{V}\right)$ of the correct transcription, where

- t is the correct message transcription
- $\widehat{\boldsymbol{F}}_y$ are the estimated features of the noisy signal
- c are modification parameters
- $\mathcal{V}$ is a clean speech model from an HMM-based ASR.

A theoretically-equivalent form of $p\left(t \mid \widehat{\boldsymbol{F}}_y(c), \mathcal{V}\right)$ is:

$$\mathcal{O}(\boldsymbol{c}) = \log\left(p\left(\widehat{\boldsymbol{F}}_y(\boldsymbol{c}) \mid \boldsymbol{t}, \mathcal{V}\right)\right) - \log\left(\sum_{\tau,\tau\neq t} p\left(\widehat{\boldsymbol{F}}_y(\boldsymbol{c}) \mid \boldsymbol{\tau}, \mathcal{V}\right) p\left(\boldsymbol{\tau} \mid \mathcal{V}\right)\right), \quad (1)$$

where τ is an index over all possible transcriptions,

$$p\left(\widehat{\boldsymbol{F}}_y(\boldsymbol{c}), \boldsymbol{s} \mid \boldsymbol{\tau}, \mathcal{V}\right) = \prod_{j=1}^{J} p\left(\widehat{\boldsymbol{f}}_y^j(\boldsymbol{c}) \mid \boldsymbol{s}^j, \mathcal{V}\right) p\left(\boldsymbol{s}^j \mid \boldsymbol{s}^{j-1}, \boldsymbol{\tau}, \mathcal{V}\right) \quad (2)$$

for some state sequence s and

$$p\left(\widehat{\boldsymbol{F}}_y(\boldsymbol{c}) \mid \boldsymbol{\tau}, \mathcal{V}\right) = \sum_{s} p\left(\widehat{\boldsymbol{F}}_y(\boldsymbol{c}), \boldsymbol{s} \mid \boldsymbol{\tau}, \mathcal{V}\right). \quad (3)$$

Phone-duration compensation: the duration of a phone is not representative of its contribution to intelligibility at the word level. Duration-invariance can be introduced through phone duration normalization.

Practical considerations: the optimization of (1) is computationally demanding when working with context-dependent speech models from ASR. Here we focus on the first term of (1) and evaluate the performance of an approximation to the desired discriminative measure.

Optimization problem:

$$\boldsymbol{c} = \underset{\boldsymbol{c}'}{\operatorname{argmax}} \sum_{l=1}^{L} \sum_{j=1}^{J_l} J_l^{-1} \log\left(p\left(\widehat{\boldsymbol{f}}_y^j(\boldsymbol{c}') \mid \boldsymbol{s}^j, \mathcal{V}\right) p\left(\boldsymbol{s}^j \mid \boldsymbol{s}^{j-1}, \boldsymbol{t}^l, \mathcal{V}\right)\right). \quad (4)$$

Modifications: i) spectral-band gain and ii) phone-energy gain adjustment (both accommodate linear energy-preservation constraints).

System Architecture

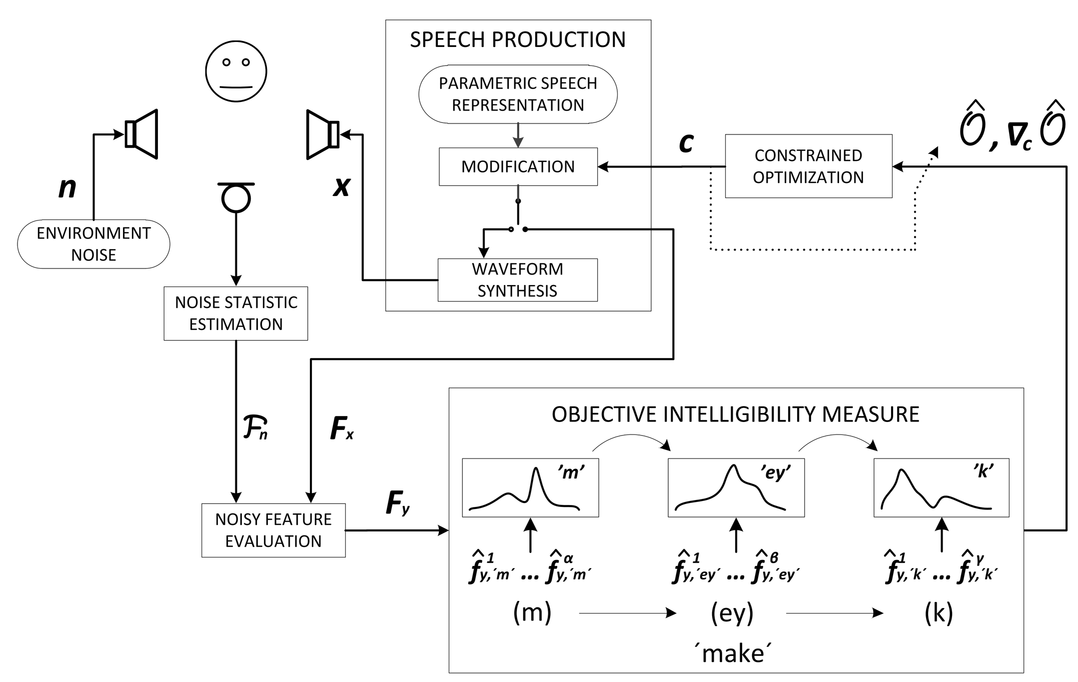

Figure 2: Diagram of the system operating on a single word.

Results

Experiment I

- Eight-band spectral gain mod. (word level/energy preserving);
- No phone-duration normalization;
- Multi-speaker babble noise at $-3\,\mathrm{dB}$, 30 sentences, 8 subjects.
- Utterance-level word recognition: $r_n = 0.38$ (natural speech) $r_m = 0.59$ (modified speech)
- High improvement significance.

Experiment II

- 50-band spectral gain mod. (word level/energy preserving);
- Phone-gain mod. (word level/energy preserving) follows spectral mod.;
- Two noise types at two SNRs, 120 sentences, 12 subjects.

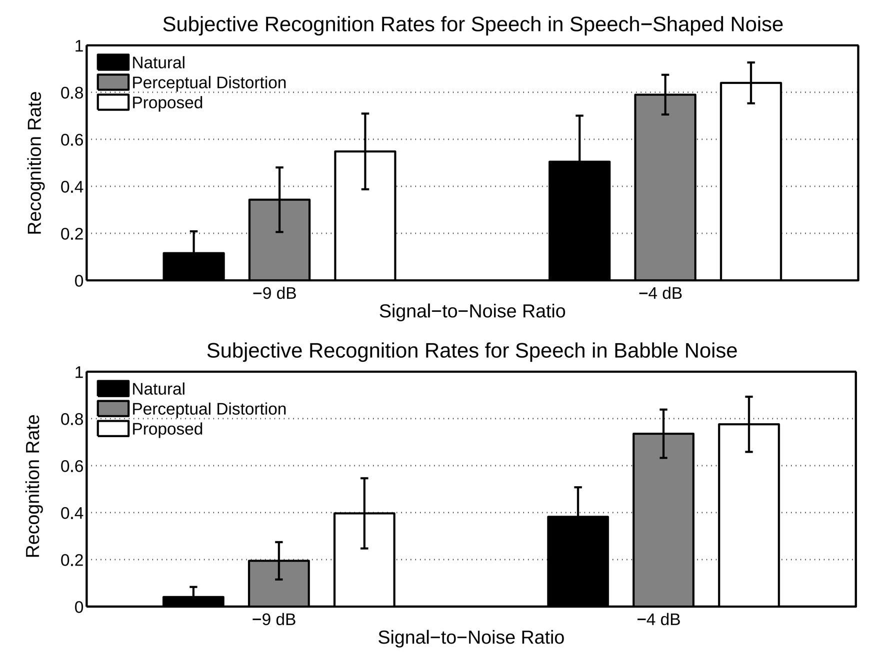

Figure 3: Results from subjective evaluation using 12 sets from the Harvard database and 12 subjects. (ref. method: Taal et al., 'A Speech Preprocessing Strategy for Intelligibility Improvement in Noise Based on a Perceptual Distortion Measure', ICASSP 2012.)

System Behaviour

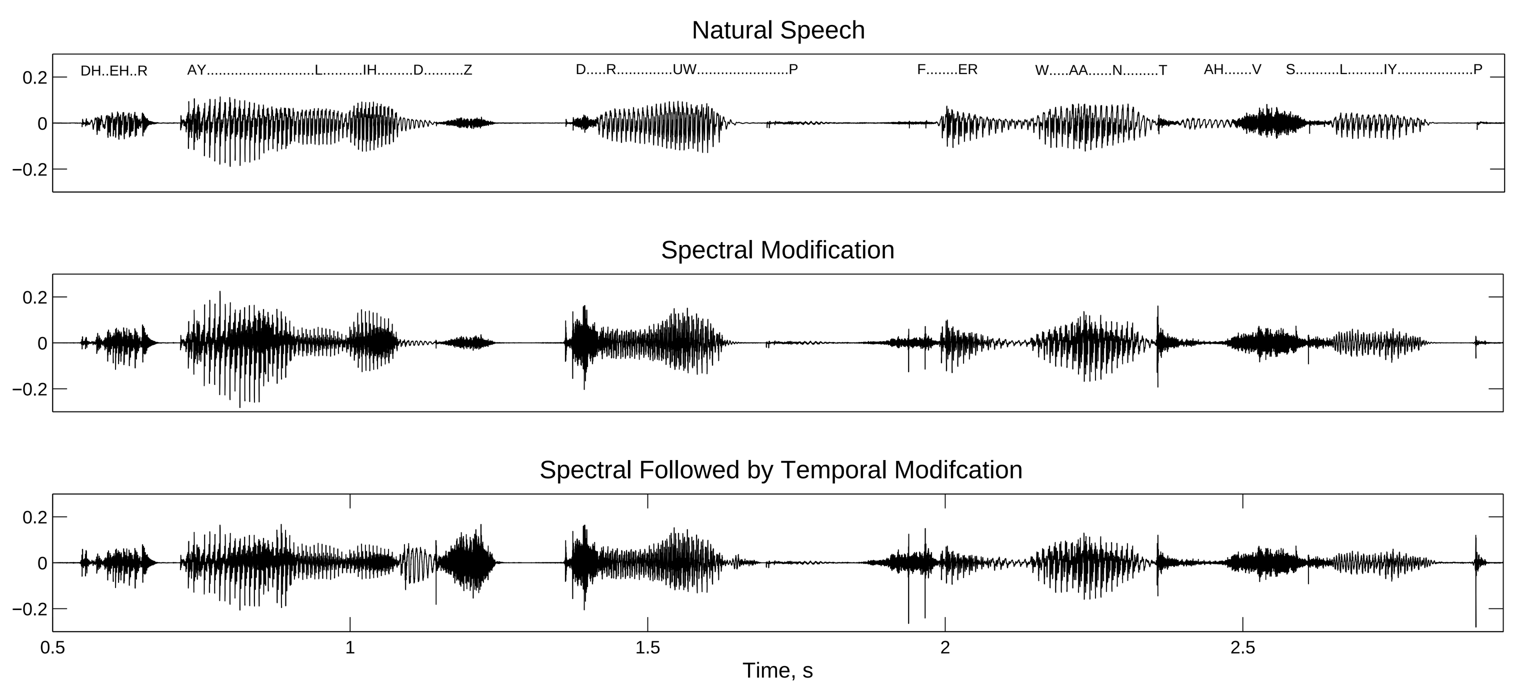

Figure 4: Effect of the adopted modifications in $-4\,\mathrm{dB\,SNR}$ speech-shaped noise.

Future Work

- Phone-level spectral modifications (sound-specific modifications);
- Evaluate the performance of the discriminative measure;
- Extend application domain: TTS, live speech, accent adaptation.