Detection of Spoken Words in Noise: Comparison of Human Performance to Maximum Likelihood Detection

Mohsen Zareian Jahromi, Jan Østergaard, Jesper Jensen

AALBORG UNIVERSITY
DENMARK

Aim and Motivation

- Here, we aim at assessing this question: **Is the human auditory system optimal in any sense**?
- Motivated by the DANTALE II test paradigm (used to evaluate the intelligibility of noisy speech by exposing human listeners to a selection of constructed noisy sentences)

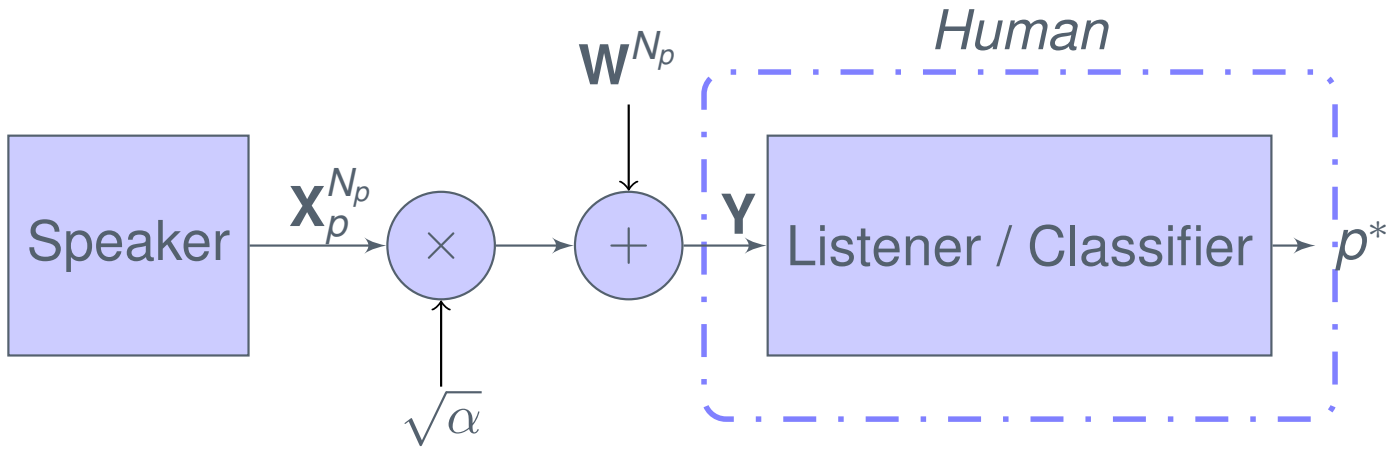
Approach

- We propose a simple model for the communication and classification of noisy speech that takes place in the DANTALE II test.
- We derive the optimal classifiers for the proposed model.
- Compare the performance of the optimal classifiers to the human performance.

The DANTALE II test paradigm

- The Danish sentence test DANTALE II has been developed to determine the speech reception threshold (SRT) in noise, i.e. the signal-to-noise ratio (SNR) for which 50% of the words can be recognized correctly. The DANTALE II database contains 150 sentences. Each sentence consists of five words from five categories (name, verb, numeral, adjective, object).
- The sentences are contaminated with additive stationary Gaussian noise with the same long-term spectrum as the sentences. The noisy signals at different SNRs are presented to the normal-hearing subjects by headphones. The subject's task is to repeat the words they heard, and the number of correct words are collected for each presented sentence. Before conducting the experiment, the subjects go through a training phase where the subject listens to versions of the noisy sentences.

Proposed Model



The model consists of three blocks:

1. Stimulus generation: a codebook of clean sentences (words), which are randomly (uniformly) selected from the DANTALE II database.
2. Communication: a communication channel with fading and additive noise.
3. The classifier is optimal in the sense of maximum a posteriori probability estimation.

Proposed Model

Assumptions on the proposed model:

1. Subjects are able to learn and memorize the words waveforms of noise-free DANTALE II sentences through the training phase.
2. Subjects are able to learn noise properties e.g. the covariance matrix of the noise.
3. When listening to the noisy sentences, the subjects do not know the SNR a priori.
4. Subject can not distinguish between different waveform realizations of the same word.
5. Subjects try to maximize the probability of correct decision.

Optimal Classifiers

The classifier or listener chooses which sentence was spoken. The optimal classifiers make a decision based on *posterior probabilities* defined as:

$$P(\mathbf{X}_p \text{ was sent } | \mathbf{Y} \text{ was received}), \tag{1}$$

where $P(\mathbf{X}_p|\mathbf{Y})$ is the conditional probability mass function (PMF) of $\mathbf{X}_p$, given $\mathbf{Y}$.

Optimal Classifiers

Optimal Bayesian Classifier

The Bayesian classifier selects the spoken sentence $\mathbf{X}_{p^*}$ maximizing the posterior probabilities:

$$p^* = \underset{p \in \{1,\ldots,M\}}{\operatorname{argmax}} \{P(\mathbf{X}_p|\mathbf{Y})\}. \tag{2}$$

Optimal Classifiers

Approximated Bayesian Classifier (continuous α)

One may argue that subjects are able to identify the SNR after having listened to a particular test stimulus, and thereby the scale factor α, before choosing the sentence (or word). In this case, we should maximise $f(\mathbf{X}_p, \alpha|\mathbf{Y})$ rather than $P(\mathbf{X}_p|\mathbf{Y})$, where $f(\mathbf{X}_p, \alpha|\mathbf{Y})$ is the conditional joint probability density function (PDF) of $\mathbf{X}_p$ and α, given $\mathbf{Y}$. This leads to the following optimisation problem:

$$(p^*, \alpha^*) = \underset{p\in\{1,\ldots,M\},\alpha\in[a,b]}{\operatorname{argmax}} \{f(\mathbf{X}_p, \alpha|\mathbf{Y})\}. \tag{3}$$

Optimal Classifiers

Approximated Bayesian Classifier (discrete α)

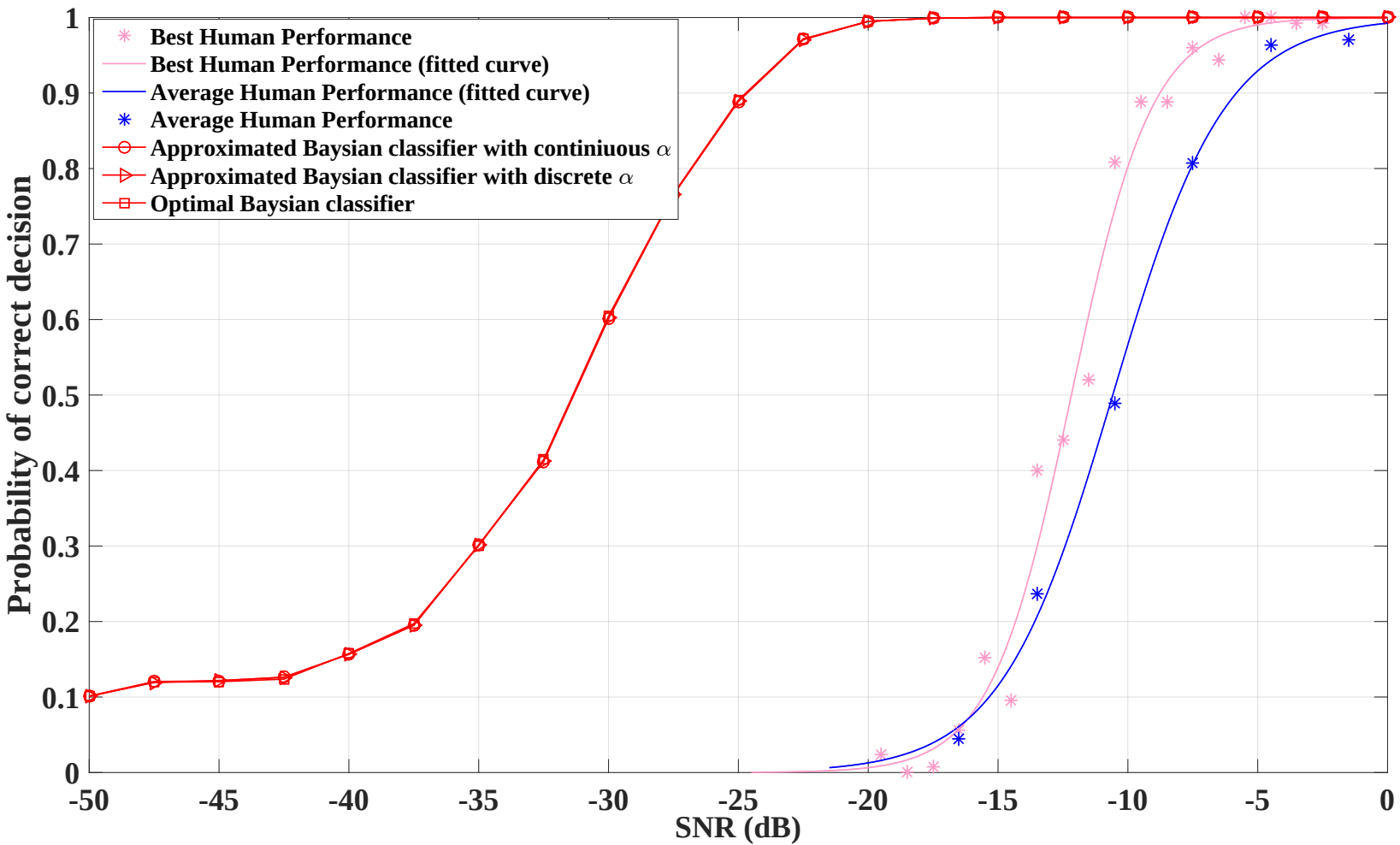
In the DANTALE II listening test, several different SNRs are used and it could be reasonable to assume that the subjects can learn these SNRs through the training phase. In this case, the scale factor is a discrete random variable ($\alpha_i, \quad i \in \{1, ..., K\}$) rather than a continuous one. Thus, we maximise $P(\mathbf{X}_p, \alpha_i|\mathbf{Y})$, where $P(\mathbf{X}_p, \alpha_i|\mathbf{Y})$ is the PMF of $\mathbf{X}_p$ and α_i, given $\mathbf{Y}$:

$$(p^*, i^*) = \underset{p\in\{1,...,M\}, i\in\{1,...,K\}}{\operatorname{argmax}} \{P(\mathbf{X}_p, \alpha_i|\mathbf{Y})\}. \tag{4}$$

Simulation Study

- The optimal classifiers are applied on each word of the Dantale II sentences.
- Results for average human performance (AHP) is obtained from where ten subjects participated in the listening test.
- Results also for best human performance (BHP) is obtained by a highly trained subject.

Simulation Study



Conclusion

- The performance of the optimal classifiers and of the human both converge to 1 for high SNRs.
- The performance of the optimal classifiers converges to 0.1 when at high noise levels (low SNRs), it classifies words randomly from 10 possible choices.
- From these results, the superior performance of the optimal classifiers is obvious, suggesting that in this very specialized task, **the human auditory system is not optimal**.

Thank you for your attention!

AALBORG UNIVERSITY
DENMARK