Humans do not maximize the probability of correct decision when recognizing DANTALE words in noise

Jahromi, Mohsen Zareian; Østergaard, Jan; Jensen, Jesper

Published in:
Proc. Interspeech 2017

DOI (link to publication from Publisher):
10.21437/Interspeech.2017-1158

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Humans do not maximize the probability of correct decision when recognizing DANTALE words in noise

Mohsen Zareian Jahromi[⋆], *Jan Østergaard*[⋆], *Jesper Jensen*[⋆†]

[⋆]Dept. of Electronic Systems, Aalborg University, Fredrik Bajers Vej 7, 9220 Aalborg, Denmark
[†]Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark
{mzj,jo,jje}@es.aau.dk

Abstract

Inspired by the DANTALE II listening test paradigm, which is used for determining the intelligibility of noisy speech, we assess the hypothesis that humans maximize the probability of correct decision when recognizing words contaminated by additive Gaussian, speech-shaped noise. We first propose a statistical Gaussian communication and classification scenario, where word models are built from short term spectra of human speech, and optimal classifiers in the sense of maximum a posteriori estimation are derived. Then, we perform a listening test, where the participants are instructed to make their best guess of words contaminated with speech-shaped Gaussian noise. Comparing the human's performance to that of the optimal classifier reveals that at high SNR, humans perform comparable to the optimal classifier. However, at low SNR, the human performance is inferior to that of the optimal classifier. This shows that, at least in this specialized task, humans are generally not able to maximize the probability of correct decision, when recognizing words.

1. Introduction

For many years, automatic speech recognition (ASR) systems have been inspired by models of the human auditory system and speech perception [1, 2, 3]. Numerous theories have been proposed of how the brain functions in response to auditory stimuli (e.g. speech signals). In [4], it is hypothesised that human listeners behave like optimal Bayesian observers, suggesting that the sensory information is re-presented probabilistically in the brain. In [5], Barlow proposes the principle of efficient coding which suggests that the brain tries to maximize the mutual information between the sensorium and its internal representation, under constraints on the efficiency of those representations. Recently Friston et al. [6, 7] proposed a free-energy principle suggesting that the brain functions in a way that it optimizes the free energy of sensations and the representation of their causes. The free energy measures the difference between the probability distribution of environmental quantities that act on the system and an arbitrary distribution encoded by its configuration. The system can minimize its free energy by changing its configuration to affect the way it samples the environment or change the distribution it encodes.

While the above works address how the brain functions in response to acoustic stimuli, they do not explain how humans recognize words in noise. In this paper, we investigate the hypothesis that listeners maximize the probability of correct decision when recognizing noisy words. To do so, we propose a communication model which reflects key properties of the DANTALE II intelligibility test [8]. In this listening test, speech intelligibility is determined by presenting speech stimuli contaminated by noise to test subjects, and recording the fraction of words understood. We then derive optimal classifiers based on the model and compare the performance of the classifiers to that of humans. Obviously, the performance of the model relies strongly on assumption about how internal representations of words and noise are stored in the brain.

In our recent work [9], we assumed for mathematic convenience that humans are able to store the temporal waveforms of each word and use those waveforms for word recognition. It is, however, well known that inner hair cells in cochlea are unable to phase lock to the signal waveforms of frequencies beyond 1.0-1.5 kHz [10, 11] and, hence, tend to represent (transmit) the overall signal power rather than temporal details. Indeed, recent work [12] suggests that humans might build internal statistical models of the words based on characteristics of the spectral contents of sub words. Inspired by this, we assume that humans are able to learn and store the power spectral density of short segments of the words. Based on these assumptions, an optimal classifier is derived in the sense of maximum a posteriori probability estimation, and its performance for speech sentences contaminated by additive speech shaped Gaussian noise is analyzed and compared to the performance of humans. Results show that humans perform comparable to that of the classifier at high SNRs, but the performance of the classifier is superior at low SNRs. This indicates that humans generally do not maximize the probability of correct decision when recognizing Dantale words in additive Gaussian, speech-shaped noise.

2. Modelling and classification

2.1. The DANTALE II test paradigm

The Danish sentence test DANTALE II [8] has been designed in order to determine the speech reception threshold (SRT), i.e. the signal-to-noise ratio (SNR) for which, e.g. 50%, of the words can be recognized correctly. The DANTALE II database contains 150 sentences sampled at 20 kHz and with a resolution of 16 bits. Each sentence is composed of five words from five categories (name, verb, numeral, adjective, object). There are 10 different words in each of the five categories. In a particular realization of the test considered in this paper, the sentences are contaminated with additive stationary Gaussian noise with the same long-term spectrum as the sentences. Conducting the test consists of two phases. In the first phase (training phase), normal-hearing subjects listen to versions of the noisy sentences to familiarize themselves with the sentence material and the noise type. In the next phase (the test phase), the listeners are exposed to the noisy sentences at different SNRs by headphones. The subject's task is to repeat the words they hear, and the number of correct words are collected for each presented sentence.

http://dx.doi.org/10.21437/Interspeech.2017-1158

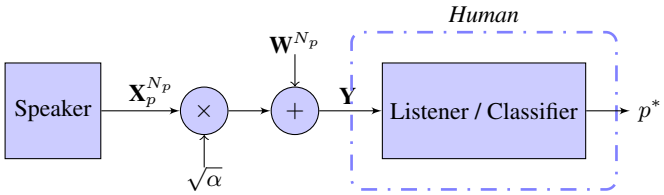

Figure 1: *The word $\boldsymbol{X}_p^{N_p}$ is conveyed over a noisy acoustic channel, and $\boldsymbol{Y}$ is received. The classifier determines which word was spoken.*

2.2. Proposed communication model

In order to be able to treat recognition of speech as a classification problem, a simple model (Fig. 1) for the DANTALE II test paradigm is proposed. The model consists of three sections:

(i) Stimulus generation: a codebook of clean words, randomly (uniformly) selected from the DANTALE II database. ("Speaker" in Fig. 1)

(ii) Communication: a communication channel with a scale factor and additive noise. The scale factor serves to modify the SNR and is randomly (uniformly) chosen from a fixed set of SNR levels. The additive noise is zero-mean coloured Gaussian and has a long term spectrum similar to the average long term spectrum of the clean speech test sentences. (Middle part of Fig. 1)

(iii) Classification: a classifier which is optimal in the sense of maximum a posteriori probability estimation.("Human" in Fig. 1)

In this work, we make two important assumptions with respect to humans ability to learn, memorize and classify words.

(i) We assume that test subjects are able to learn and store a model of the words based on the spectral envelope contents of sub words encountered during the training phase. In a similar manner, we assume that subjects create an internal noise model. In our classifier, this is achieved by allowing the classifier access to training data in terms of average short-term speech spectra of the clean speech and of the noise.

(ii) Subjects are instructed to make a best guess of each noisy word. We reflect this by designing a classifier which maximizes the probability of correct word detection.

In addition, we impose the following requirements on the subjective listening test and the classifier:

(i) When listening to the stimuli, i.e. the noisy sentences (words), the subjects are not informed about the SNRs a priori. In a similar manner, the classifier does not rely on a priori SNR knowledge.

(ii) We assume that subjects do not know when the words (noisy stimuli) start a priori.[1] Similarly, the classifier has no a priori information about the locations of the words within the noisy sentences.

2.3. Signal Model

The DANTALE II stimulus generation/selection process for the particular test with Gaussian speech shaped noise, considered in this paper, is modelled as follows: A word from a fixed dictionary containing M different words is selected. A real-valued vector $\mathbf{X}_p^{N_p} = [x_p(1), x_p(2), ..., x_p(N_p)]^T, p \in \{1, ..., M\}$, represents the time domain waveform of the p^{th} word consisting of N_p samples. The word $\mathbf{X}_p^{N_p}$ is chosen uniformly from the dictionary. The waveform is multiplied by a scale factor controlling the SNR. The scale factor α is also uniformly selected $\alpha \sim \mathcal{U}(a, b)$ where $b > a > 0$. The Gaussian speech shaped noise waveform $\mathbf{W}^{N_p}$ is given by $\mathbf{W}^{N_p} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_w})$. The received word $\mathbf{Y}$ is a random vector expressed as:

$$\mathbf{Y} = \sqrt{\alpha}\mathbf{X}_p^{N_p} + \mathbf{W}^{N_p}. \tag{1}$$

Here, we first assume that all words have the same length ($N_p = N, p = 1, ..., M$), so we ignore superscript N_p. We segment each word, $\mathbf{X}_p$, into L small frames and assume that each frame is weak-sense stationary and can be modeled by a zero-mean autoregressive Gaussian process of order n, e.g. [13]:

$$f_{\mathbf{X}_{p,z}}(\mathbf{X}_{p,z}) = |2\pi\mathbf{\Sigma}_{\mathbf{x}_{p,z}}|^{-\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{X}_{p,z}^T\mathbf{\Sigma}_{\mathbf{x}_{p,z}}^{-1}\mathbf{X}_{p,z}), \tag{2}$$

where $f_{\mathbf{X}_{p,z}}$ is the probability density function (PDF) of the z^{th} frame of the p^{th} word, and $\mathbf{\Sigma}_{\mathbf{x}_{p,z}}$ is the covariance of the z^{th} frame of that word (the length of each frame is K). If $\mathbf{a}_{\mathbf{x}_{p,z}} = [1, a_{\mathbf{x}_{p,z}}(1), ..., a_{\mathbf{x}_{p,z}}(n)]^T$ denotes the vector of the linear predictive (LP) coefficients of $\mathbf{X}_{p,z}$, and $\sigma^2_{\mathbf{x}_{p,z}}$ is the variance of the prediction error respectively, the covariance matrix $\mathbf{\Sigma}_{\mathbf{x}_{p,z}}$ is obtained as [13]

$$\mathbf{\Sigma}_{\mathbf{x}_{p,z}} = \sigma^2_{\mathbf{x}_{p,z}}(A^T_{\mathbf{x}_{p,z}}A_{\mathbf{x}_{p,z}})^{-1}, \tag{3}$$

where $A_{\mathbf{x}_{p,z}}$ is the $K \times K$ lower triangular Toeplitz matrix with $[1, a_{\mathbf{x}_{p,z}}(1), ..., a_{\mathbf{x}_{p,z}}(n), 0, 0, 0]^T$ as its first column. We assume that all frames are mutually independent, which implies that $f_{\mathbf{X}_p}(\mathbf{X}_p)$ can be written as:

$$f_{\mathbf{X}_p}(\mathbf{X}_p) = \prod_{z=1}^{L} f_{\mathbf{X}_{p,z}}(\mathbf{X}_{p,z}) = \left(\prod_{z=1}^{L} |2\pi\mathbf{\Sigma}_{\mathbf{x}_{p,z}}|^{-\frac{1}{2}}\right) \exp\left(-\frac{1}{2}\sum_{z=1}^{L}\mathbf{X}_{p,z}^T\mathbf{\Sigma}_{\mathbf{x}_{p,z}}^{-1}\mathbf{X}_{p,z}\right).$$

2.4. Optimal Classifier

2.4.1. Optimal Bayesian Classifier

The classifier chooses which word was spoken. The classifier makes a decision by maximizing the *posterior probability*:

$$P(\mathbf{X}_p \text{ was sent } |\mathbf{Y} \text{ was received}), \tag{4}$$

where $P(\mathbf{X}_p|\mathbf{Y})$ is the conditional probability mass function (PMF) of $\mathbf{X}_p$, given $\mathbf{Y}$. The classifier selects the spoken word $\mathbf{X}_{p^*}$ maximizing the posterior probabilities:

$$p^* = \underset{p\in\{1,...,M\}}{\operatorname{argmax}} \{P(\mathbf{X}_p|\mathbf{Y})\}. \tag{5}$$

Lemma 1. *The optimal p^*, maximizing* (5) *is given by (see Appendix A.1 for the proof):*

$$p^* = \underset{p\in\{1,...,M\}}{\operatorname{argmax}} \int_a^b \left(\prod_{z=1}^{L} |2\pi\mathbf{\Sigma}_{\mathbf{Y}_z}|^{-\frac{1}{2}}\right) \exp(\frac{\sum_{z=1}^{L}\mathbf{Y}_z^T\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}\mathbf{Y}_z}{-2})d\alpha,$$

[1] In the DANTALE II listening test, the stimuli consist of a noise-ramp-up, the noisy speech, and then a noise-ramp-down.

where $\mathbf{Y}_z$ is the z^{th} frame of the received word. $\mathbf{\Sigma}_{\mathbf{Y}_z} = \alpha\mathbf{\Sigma}_{\mathbf{x}_{p,z}} + \mathbf{\Sigma}_{\boldsymbol{w}}$ is the covariance matrix of that frame, and L is the number of frames.

2.4.2. Approximate Bayesian Classifier (continuous α)

One may argue that subjects are able to identify the SNR and thereby the scale factor α after having listened to a particular test stimulus, before deciding on the 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:

$$(p^*, \alpha^*) = \underset{p\in\{1,\dots,M\},\alpha\in[a,b]}{\operatorname{argmax}} \{f(\mathbf{X}_p, \alpha|\mathbf{Y})\}. \quad (6)$$

Lemma 2. *The optimal pair (p^*, α^*), maximizing (6) is given by :*[2]*(see Appendix A.2 for the proof)*

$$p^* = \underset{p\in\{1,\dots,M\}}{\operatorname{argmax}} \left\{ -\sum_{z=1}^{L} \left(\mathbf{Y}_z^T (\mathbf{\Sigma}^*_{\mathbf{Y}_z})^{-1} \mathbf{Y}_z + \log|\mathbf{\Sigma}^*_{\mathbf{Y}_z}| \right) \right\},$$

where $\mathbf{\Sigma}^*_{\mathbf{Y}_z} = \alpha^*\mathbf{\Sigma}_{\mathbf{x}_{p,z}} + \mathbf{\Sigma}_{\boldsymbol{w}}$ and α^* is obtained by solving the following equation with respect to α:

$$\sum_{z=1}^{L} \left(-\mathbf{Y}_z^T \left(\left(\alpha\mathbf{\Sigma}_{\mathbf{x}_{p,z}} + \mathbf{\Sigma}_{\boldsymbol{w}}\right)^{-1} \mathbf{\Sigma}_{\mathbf{x}_{p,z}} \left(\alpha\mathbf{\Sigma}_{\mathbf{x}_{p,z}} + \mathbf{\Sigma}_{\boldsymbol{w}}\right)^{-1} \right) \mathbf{Y}_z \right.$$
$$\left. + \operatorname{tr}\left(\left(\alpha\mathbf{\Sigma}_{\mathbf{x}_{p,z}} + \mathbf{\Sigma}_{\boldsymbol{w}}\right)^{-1} \mathbf{\Sigma}_{\mathbf{x}_{p,z}} \right) \right) = 0.$$

2.4.3. Approximate Bayesian Classifier (discrete α)

In the version of the DANTALE II listening test considered in this paper, a fixed limited set of SNRs are used, and it could be reasonable to assume that the subjects can identify all these SNRs through the training phase. In this case, the scale factor is a discrete random variable ($\alpha_i, \quad i \in \{1, ..., S\}$) 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 $\mathbf{Y}$. The optimization problem in (6) can be rewritten by:

$$(p^*, i^*) = \underset{p\in\{1,\dots,M\},i\in\{1,\dots,S\}}{\operatorname{argmax}} \{P(\mathbf{X}_p, \alpha_i|\mathbf{Y})\}. \quad (7)$$

Lemma 3. *The optimal pair (p^*, i^*), maximizing (7) is given by:*

$$(p^*, i^*) = \underset{p\in\{1,\dots,M\},i\in\{1,\dots,S\}}{\operatorname{argmax}} -\sum_{z=1}^{L} \left(\mathbf{Y}_z^T (\mathbf{\Sigma}^i_{\mathbf{Y}_z})^{-1} \mathbf{Y}_z + \log|\mathbf{\Sigma}^i_{\mathbf{Y}_z}| \right),$$

where $\mathbf{\Sigma}^i_{\mathbf{Y}_z} = \alpha_i\mathbf{\Sigma}_{\mathbf{x}_{p,z}} + \mathbf{\Sigma}_{\boldsymbol{w}}$.

The proof of Lemma 3 mostly follows the proof of Lemma 2, and has been eliminated due to space limitations.

2.5. Temporal Misalignment

In order to take into account the requirement (ii) that subjects do not know when the sentence (word) starts a priori, a window with the same size as the word is shifted within the stimuli and for each shift, the likelihoods $P(\mathbf{X}_p|\mathbf{Y}^w)$ for the optimal Bayesian classifier, $P(\mathbf{X}_p, \alpha|\mathbf{Y}^w)$ for the approximated Bayesian classifier when α is continuous, and $P(\mathbf{X}_p, \alpha_i|\mathbf{Y}^w)$ for the approximated Bayesian classifier when α is discrete are calculated. $\mathbf{Y}^w$ denotes the portion of $\mathbf{Y}$ for each shift w. Finally p^* is given by:

$$p^* = \underset{p\in\{1,\dots,M\}}{\operatorname{argmax}} \{\max_w\{P(\mathbf{X}_p|\mathbf{Y}^w)\}\},$$
$$(p^*, \alpha^*) = \underset{p\in\{1,\dots,M\},\alpha\in[a,b]}{\operatorname{argmax}} \{\max_w\{P(\mathbf{X}_p, \alpha|\mathbf{Y}^w)\}\},$$
$$(p^*, i^*) = \underset{p\in\{1,\dots,M\},i\in\{1,\dots,S\}}{\operatorname{argmax}} \{\max_w\{P(\mathbf{X}_p, \alpha_i|\mathbf{Y}^w)\}\}.$$

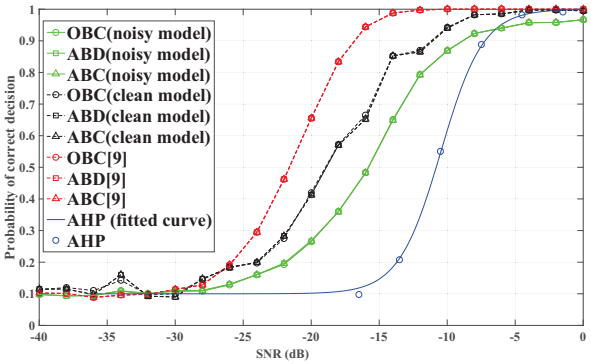

Figure 2: *Comparison of human's performance with the optimal classifiers for word detection. (OBC: Optimal Bayesian classifier, ABD: Approximate Bayesian classifier with discrete α, ABC: Approximate Bayesian classifier with continuous α, AHP: average human performance).*

3. Simulation Study

In the DANTALE II database, each word ($M = 10$) has 15 different realizations ($\mathbf{X}_p^j, j \in \{1, 2, ..., 15\}$, where subscript j denotes the j^{th} realization of the p^{th} word). In our simulations, we use 14 different realizations of words for training (building covariance matrices), and one realization of the words for testing. In this way, we assume that the listeners learn one statistical model (covariance matrices) of sub words for all realizations of that word through the training phase. To build the covariance matrix for each frame, after segmenting each word into L non overlapping frames (the size of each frame is 20 ms and $L = 20$), a long vector containing the same frame of 14 realizations is created $[\mathbf{X}^1_{p,z}, \mathbf{X}^2_{p,z}, ..., \mathbf{X}^{14}_{p,z}]$. Then the vector of the LP coefficients ($\mathbf{a}_{\mathbf{x}_{p,z}}$) of this long vector is obtained, and finally the covariance matrix of this frame ($\mathbf{\Sigma}_{\mathbf{x}_{p,z}}$) is calculated using (3). We refer to this model as "noisy model". Fig. 2 demonstrates the performance of the classifiers as a function of SNR. Using the leave-one-out method, where 14 realizations are used for training and the last one for testing, we obtain 15 results whose average is used as the final result (green solid curves). It might also be reasonable to assume that subjects are able to store one statistical model for each realization of a word. In this model, the covariance matrix for the z^{th} frame of the j^{th} realization of the p^{th} word ($\mathbf{\Sigma}^j_{\mathbf{x}_{p,z}}$) is obtained based on LP coefficients derived from frame $\mathbf{X}^j_{p,z}$. The results for this model ("clean model") is shown by dashed solid curves.

3.1. Human performance

In order to measure the human performance, we performed a listening test where the DANTALE II material was used. 18 normal-hearing subjects participated in this test. They were

[2] We assume that $a \leq \alpha^* \leq b$, otherwise the nearest point should be chosen.

presented to the DANTALE words contaminated by speech-shaped noise by headphones and they chose the words they heard using a GUI interface[3]. In the training phase, the subjects were exposed to 12 noisy sentences at 6 different SNRs, where each SNR was repeated twice. The sentences in the training phase are also composed from the DANTALE II data base. In the test phase, each subject listened to 48 sentences (6 SNRs $\times$ 8 repetitions). The average result for this listening test is shown in Fig. 2 (abbreviated as AHP). The fitted line is a ML (Maximum Likelihood)-fitted logistic function of the form $f(x) = \frac{1-\frac{1}{10}}{1+\exp(cx+d)} + \frac{1}{10}$. The 50% SRT is approximately -10 dB which is well in line with similar tests performed in literature [14] for this noise type.

4. Discussion

As seen in Fig. 2, all three classifiers (OBC, ABC, ABD) perform very similar on all tasks. Thus, in this test the alphabet of the SNR and prior assumptions on it is insignificant. For low SNRs, the performance of all optimal classifiers converges to 0.1. This is because at high noise levels, the classifiers choose words randomly (from $M = 10$ words). Fig. 2 also shows the performance of the optimal classifiers (red dashed curves) for the deterministic model [9] (which assumes that classification can be made based on knowledge of word *waveforms*). We observe that the classifiers performance when relying on statistical models (i.e. noise and speech covariances) is closer to the performance of humans. This is in line with results from [12], where the performance of a simple hidden markov model (HMM)-based recognition system (trained on DANTALE II database) is *similar* to that of humans. It can be seen that the optimal classifiers both based on the statistical models and the deterministic model outperform the humans' performance at low SNRs and coincide with the humans' performance at high SNRs. It is worth mentioning that the optimal classifiers are based on potentially debatable statistical assumptions with regard to speech signals (e.g. Gaussian samples and statistically independent sub-words). However, if a truly optimal classifier was used, then the performance of the classifiers ("noisy" and "clean" models) would improve further. This suggests that under certain assumptions, humans do not maximize the probability of correct decision when recognizing DANTALE words in additive Gaussian, speech-shaped noise.

A. Proof of lemmas

A.1. Proof of lemma 1

Using Bayes' theorem, (4) can be written as [15]:

$$P(\mathbf{X}_p|\mathbf{Y}) = \frac{f_{\mathbf{Y}|\mathbf{X}_p}(\mathbf{Y}|\mathbf{X}_p)P(\mathbf{X}_p)}{f_{\mathbf{Y}}(\mathbf{Y})}, \tag{8}$$

where $f_{\mathbf{Y}|\mathbf{X}_p}$ is the conditional PDF of the received word, given word $\mathbf{X}_p$, $f_{\mathbf{Y}}$ is the PDF of the received word, and $P(\mathbf{X}_p)$ is the probability that $\mathbf{X}_p$ is spoken. Since $P(\mathbf{X}_p) = \frac{1}{M}, \forall p$, and $f_{\mathbf{Y}}(\mathbf{Y})$ is independent of $\mathbf{X}_p$, from (8), (5) can be rewritten:

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

[3]The authors would like to thank Asger Heidemann Andersen for making the interface available to us.

If the received word is divided into L small frames, from (2), it is easy to show that

$$f_{\mathbf{Y}_z|\mathbf{X}_p,\alpha}(\mathbf{Y}_z|\mathbf{X}_p,\alpha) = |2\pi\mathbf{\Sigma}_{\mathbf{Y}_z}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{Y}_z^T\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}\mathbf{Y}_z\right),$$

where $f_{\mathbf{Y}_z|\mathbf{X}_p,\alpha}(\mathbf{Y}_z|\mathbf{X}_p,\alpha)$ is the conditional PDF of the z^{th} frame of the received word, given $\mathbf{X}_p$ and α, and $\mathbf{\Sigma}_{\mathbf{Y}_z} = \alpha\mathbf{\Sigma}_{\mathbf{x}_{p,z}} + \mathbf{\Sigma}_{\mathbf{w}}$. With the assumption that frames of the received word are mutually independent, we get:

$$\begin{aligned} f_{\mathbf{Y}|\mathbf{X}_p,\alpha}(\mathbf{Y}|\mathbf{X}_p,\alpha) &= \prod_{z=1}^{L} f_{\mathbf{Y}_z|\mathbf{X}_p,\alpha}(\mathbf{Y}_z|\mathbf{X}_p,\alpha) \\ &= \left(\prod_{z=1}^{L}|2\pi\mathbf{\Sigma}_{\mathbf{Y}_z}|^{-\frac{1}{2}}\right)\exp\left(-\frac{1}{2}\sum_{z=1}^{L}\mathbf{Y}_z^T\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}\mathbf{Y}_z\right). \end{aligned} \tag{9}$$

Since $\mathbf{X}_p$ and α are mutually independent, $f_{\mathbf{Y}|\mathbf{X}_p} = \int_a^b f_{\mathbf{Y}|\mathbf{X}_p,\alpha} f_\alpha d\alpha$. Using (9), we have:

$$\begin{aligned} p^* &= \underset{p\in\{1,\ldots,M\}}{\operatorname{argmax}} \left\{\int_a^b f_{\mathbf{Y}|\mathbf{X}_p,\alpha}(\mathbf{Y}|\mathbf{X}_p,\alpha)d\alpha\right\} \\ &= \underset{p\in\{1,\ldots,M\}}{\operatorname{argmax}} \int_a^b \left(\prod_{z=1}^{L}|2\pi\mathbf{\Sigma}_{\mathbf{Y}_z}|^{-\frac{1}{2}}\right)\exp\left(\frac{\sum_{z=1}^{L}\mathbf{Y}_z^T\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}\mathbf{Y}_z}{-2}\right)d\alpha. \end{aligned}$$

A.2. Proof of lemma 2

According to Bayes' theorem, $f(\mathbf{X}_p,\alpha|\mathbf{Y})$ can be written as:

$$f(\mathbf{X}_p,\alpha|\mathbf{Y}) = \frac{f_{\mathbf{Y}|\mathbf{X}_p,\alpha}(\mathbf{Y}|\mathbf{X}_p,\alpha)f_{\mathbf{X}_p,\alpha}(\mathbf{X}_p,\alpha)}{f_{\mathbf{Y}}(\mathbf{Y})}, \tag{10}$$

where $f_{\mathbf{Y}|\mathbf{X}_p,\alpha}$ is the conditional probability density function (PDF) of the received word, given word $\mathbf{X}_p$ and α, $f_{\mathbf{Y}}$ is the PDF of the received word, and $f_{\mathbf{X}_p,\alpha}(\mathbf{X}_p,\alpha)$ is the joint PDF of $\mathbf{X}_p$ is spoken and α is used. $\mathbf{X}_p$ and α are mutually independent, so $f_{\mathbf{X}_p,\alpha}(\mathbf{X}_p,\alpha) = P(\mathbf{X}_p)f_\alpha = \frac{1}{M(b-a)}$. Using (10) and (9), (6) can be expressed as:

$$\begin{aligned} (p^*,\alpha^*) &= \underset{p\in\{1,\ldots,M\},\alpha\in[a,b]}{\operatorname{argmax}} \{f_{\mathbf{Y}|\mathbf{X}_p,\alpha}(\mathbf{Y}|\mathbf{X}_p,\alpha)\} \\ &= \underset{p\in\{1,\ldots,M\},\alpha\in[a,b]}{\operatorname{argmax}} \left(\prod_{z=1}^{L}|2\pi\mathbf{\Sigma}_{\mathbf{Y}_z}|^{-\frac{1}{2}}\right)\exp\left(-\frac{1}{2}\sum_{z=1}^{L}\mathbf{Y}_z^T\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}\mathbf{Y}_z\right). \end{aligned}$$

By applying the logarithm, we get :

$$(p^*,\alpha^*) = \underset{p\in\{1,\ldots,M\},\alpha\in[a,b]}{\operatorname{argmax}} -\sum_{z=1}^{L}\left(\mathbf{Y}_z^T\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}\mathbf{Y}_z + \log|\mathbf{\Sigma}_{\mathbf{Y}_z}|\right).$$

Above equation indicates that the decoder chooses the pair (p,α) maximizing $g = -\sum_{z=1}^{L}\left(\mathbf{Y}_z^T\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}\mathbf{Y}_z + \log|\mathbf{\Sigma}_{\mathbf{Y}_z}|\right)$. $\mathbf{Y}_z$ is assumed as a constant at the decoder, so using that $\frac{\partial\log|\mathbf{\Sigma}_{\mathbf{Y}_z}|}{\partial\alpha} = \operatorname{tr}\left(\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}\frac{\partial\mathbf{\Sigma}_{\mathbf{Y}_z}}{\partial\alpha}\right)$, and that $\frac{\partial\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}}{\partial\alpha} = -\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}\frac{\partial\mathbf{\Sigma}_{\mathbf{Y}_z}}{\partial\alpha}\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}$, we find

$$\begin{aligned} &\alpha^* \Rightarrow \frac{\partial g}{\partial\alpha} = 0 \Rightarrow \sum_{z=1}^{L}\left(-\mathbf{Y}_z^T\left(\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}\mathbf{\Sigma}_{\mathbf{x}_{p,z}}\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}\right)\mathbf{Y}_z + \right. \\ &\left.\operatorname{tr}\left(\mathbf{\Sigma}_{\mathbf{Y}_z}^{-1}\mathbf{\Sigma}_{\mathbf{x}_{p,z}}\right)\right) = 0 \\ &p^* = \underset{p\in\{1,\ldots,M\}}{\operatorname{argmax}}\left\{-\sum_{z=1}^{L}\left(\mathbf{Y}_z^T(\mathbf{\Sigma}_{\mathbf{Y}_z}^*)^{-1}\mathbf{Y}_z + \log|\mathbf{\Sigma}_{\mathbf{Y}_z}^*|\right)\right\}, \end{aligned}$$

where $\mathbf{\Sigma}_{\mathbf{Y}_z}^* = \alpha^*\mathbf{\Sigma}_{\mathbf{x}_{p,z}} + \mathbf{\Sigma}_{\mathbf{w}}$.

5. References

[1] K. Wang and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 3, pp. 421–435, 1994.

[2] R. M. Stern, "Applying physiologically-motivated models of auditory processing to automatic speech recognition," in *invited talk at the International symposium on auditory and audiological research*, 2011.

[3] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *Information Theory, IEEE Transactions on*, vol. 38, no. 2, pp. 824–839, 1992.

[4] D. C. Knill and A. Pouget, "The bayesian brain: the role of uncertainty in neural coding and computation," *Trends in neurosciences*, vol. 27, no. 12, pp. 712–719, 2004.

[5] H. B. Barlow, "Possible principles underlying the transformations of sensory messages," *Sensory Communication*, pp. 217–234, 1961.

[6] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[7] K. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain," *Journal of Physiology-Paris*, vol. 100, no. 1, pp. 70–87, 2006.

[8] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a danish sentence test in noise: Diseño, optimización y evaluación de la prueba danesa de frases en ruido," *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, 2003.

[9] M. Z. Jahromi, J. Østergaard, and J. Jensen, "Detection of spoken words in noise: Comparison of human performance to maximum likelihood detection," in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2016.

[10] J. O. Pickles, *An introduction to the physiology of hearing*. Academic press London, 1988, vol. 2.

[11] C. J. Plack, *The Sense of Hearing*. Psychology Press, 2013.

[12] M. R. Schädler, A. Warzybok, S. Hochmuth, and B. Kollmeier, "Matrix sentence intelligibility prediction using an automatic speech recognition system," *International Journal of Audiology*, vol. 54, no. sup2, pp. 100–107, 2015.

[13] S. Godsill, "The restoration of degraded audio signals." Ph.D. dissertation, University of Cambridge, 1993.

[14] A. H. Andersen, J. M. De Haan, Z. H. Tan, and J. Jensen, "A binaural short time objective intelligibility measure for noisy and enhanced speech," *Interspeech-2015*, 2015.

[15] J. G. Proakis, *Digital communications*. New York: McGraw-Hill, 1995.