Factor Graph Decoding for Speech Presence Probability Estimation

Thomas Glarner, Mohammad Mahdi Momenzadeh, Lukas Drude, Reinhold Haeb-Umbach

Department of Communications Engineering, Paderborn University, 33098 Paderborn, Germany
Email: {glarner,drude,haeb}@nt.uni-paderborn.de
Web: nt.uni-paderborn.de

Abstract

This paper is concerned with speech presence probability estimation employing an explicit model of the temporal and spectral correlations of speech. An undirected graphical model is introduced, based on a Factor Graph formulation. It is shown that this undirected model cures some of the theoretical issues of an earlier directed graphical model. Furthermore, we formulate a message passing inference scheme based on an approximate graph factorization, identify this inference scheme as a particular message passing schedule based on the turbo principle and suggest further alternative schedules. The experiments show an improved performance over speech presence probability estimation based on an IID assumption, and a slightly better performance of the turbo schedule over the alternatives.

1 Introduction

Speech Presence Probability (SPP) estimation refers to the estimation of the presence of speech in the Short Time Fourier Transform (STFT) domain at the resolution of individual time-frequency (TF) bins. SPP estimators are used, e.g., in several state-of-the-art noise tracking algorithms [1], for *a-priori* SNR estimation [2], or for mask estimation in acoustic beamforming [3] and therefore are an important component of modern speech enhancement systems.

SPP estimators exploit the correlations of the signal in time and/or frequency direction in some way or another. In a popular approach, the SPP estimation relies on a decision-directed estimate of the a priori signal-to-noise ratio [4], which in turn depends on the clean speech estimate, thus coupling SPP estimation and speech enhancement. Gerkmann et al. decouple the two and employ spectral and temporal smoothing of the a posteriori SNR and the SPP using heuristically chosen parameters [5], while Momeni et al. apply a sliding window directly to the complex-valued observations in the STFT domain to determine the statistics of the signal and noise, from which in turn the SPP is estimated [6].

Tran Vu and Haeb-Umbach propose a formulation that directly models the underlying statistical dependencies in time and frequency direction, employing a two-dimensional hidden Markov Model (2DHMM) [7]. For each TF-slot a latent binary variable is introduced which indicates presence or absence of speech. The temporal and spectral correlations of a speech signal are then captured by the transition probabilities of the 2D Markov grid, which can be learned in a training phase. An efficient inference algorithm has been developed which is an instance of the turbo principle known from coding literature [8].

In this paper we extend and generalize this approach in two important aspects. First, we propose an equivalent Undirected Graphical Model (UGM), i.e., a Markov Random Field (MRF), and show that this has important theoretical advantages over the directed HMM. While MRFs have already been used in speech enhancement for providing an MMSE estimator for the spectral amplitude [9], they are employed here for SPP estimation for the first time. Second, viewing the aforementioned turbo algorithm as a particular message passing scheme for inference in graphical models with loops, alternative schemes will be investigated, thus introducing general factor graph decoding as an approach to SPP estimation.

The remainder of the paper is organized as follows. In Section 2, the signal model is presented, and the SPP estimation problem is formulated. Section 3 introduces the UGM, a Factor Graph formulation and the inference algorithm. The UGM is contrasted with the Directed Graphical Model of [7] in Section 4, giving reasons why the UGM is to be preferred. In Section 5, the inference algorithm is generalized to different message passing schedules based on differing propagation paths through the graph, while Section 6 provides an experimental comparison of different scheduling types with respect to their performance on the SPP estimation task.

2 Problem Formulation

As usual, our signal model is formulated in the STFT domain and assumes an additive mixture of speech and noise. According to [7], a latent binary variable $Z(m,k)$ is introduced which indicates presence or absence of speech:

$$X(m,k) = \begin{cases} N(m,k) & \text{if } Z(m,k)=0, \\ S(m,k)+N(m,k) & \text{if } Z(m,k)=1, \end{cases} \quad (1)$$

where $1 \leq m \leq M$ is the STFT frame index, $1 \leq k \leq K$ is the frequency index, $S(m,k)$, $N(m,k)$, and $X(m,k)$ are the STFT Coefficients corresponding to the speech, noise, and observed microphone signal, respectively.

Following [5, 7] observations will be modeled in terms of the *a-posteriori* SNR

$$\zeta(m,k) = |X(m,k)|^2 / \hat{\Phi}_{NN}(k), \quad (2)$$

rather than of the microphone signal directly. Here, $\hat{\Phi}_{NN}(k)$ is a (coarse) estimate of the noise power spectral density. With the assumption of complex Gaussian distributed random noise, the observation likelihood is computed from an exponential distribution with a parameter depending on the latent variable:

$$p(\zeta(m,k)|Z(m,k)=i) = \lambda_i \exp(-\lambda_i \zeta(m,k)). \quad (3)$$

Here, $\lambda_0 = 1$ while λ_1 depends on the *a-priori* SNR and is set empirically.

The task of SPP estimation is the inference of the posterior probability of the latent variable:

$$\gamma(m,k) = \Pr\left(Z(m,k) \,|X(1{:}M, 1{:}K)\right). \quad (4)$$

Here, $1{:}J$ denotes that the corresponding index variable runs from 1 to J. Thus, in the above equation, the observations from all frames and frequencies are employed for the estimation of $Z(m,k)$.

ISBN 978-3-8007-4275-2

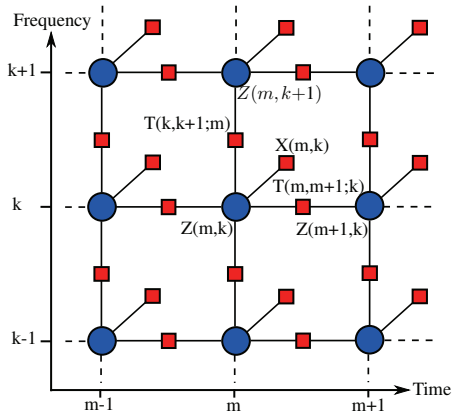


Figure 1: Part of the factor graph for the Undirected Model

It is assumed that the observations are conditionally independent if the latent variables are known. Therefore, all dependencies are modeled by the latent variables. The SPP can in theory be calculated by applying the Bayes Theorem, factorizing into the joint state probability and the observation likelihoods, and marginalizing over all state variables except $Z(m,k)$:

$$\begin{aligned}&\Pr(Z(m,k)|X(1{:}M,1{:}K)) \qquad (5)\\ &\propto \sum_{\sim Z(m,k)} \Pr(Z(1{:}M,1{:}K)) \prod_{\forall(m,k)} p(X(m,k)|Z(m,k)),\end{aligned}$$

where the operator $\sum_{\sim x}$ is taken from [10] and represents summing over all variables except x. However, a direct calculation of the marginalization is not tractable in practice as it requires 2^{MK-1} additions where MK is usually in the order of 10^5. To handle this issue, it is necessary to constrain the statistical dependencies between the latent variables.

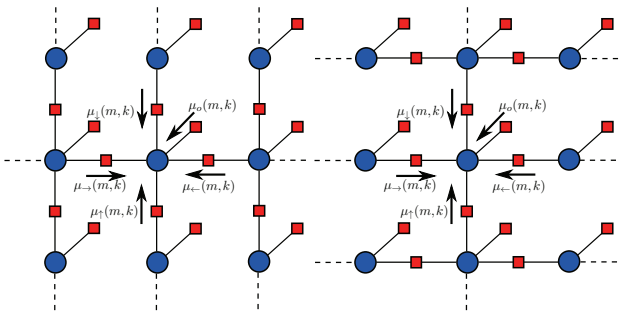
The following sections cover different formalisms to introduce reasonable constraints with the help of graphical models.

3 Undirected Graphical Model

In this Section an Undirected Graphical Model (UGM) is proposed to model temporal and spectral correlations in a speech signal. UGMs are also known as MRFs in the literature [11]. The Factor Graph framework is used [10] to provide a higher cohesion between the mathematical formulation and the graphical representation. This has the advantage of explicitly visualizing the statistical dependencies and allows for a greater flexibility in the inference scheme as explained later.

A local subsection of the graph is shown in Fig. 1. Here, the horizontal direction corresponds to the time axis, while the vertical is the frequency axis. A blue circle indicates a random variable (the speech presence $Z(m,k)$ in the corresponding TF slot). Here, the factor graph style of [10] is used where statistical dependencies are captured by the square-shaped factor nodes while the variable nodes are circular shaped nodes. Observation likelihoods are represented as factor nodes as well since the observations $X(m,k)$ are fixed. Using this graphical representation, the joint state probability is assumed to factorize as

$$\Pr(Z(1{:}M,1{:}K)) \propto \prod_{(z_i,z_j)\in\mathscr{P}(Z(1:M,1:K))} T(z_i,z_j), \qquad (6)$$



(a) Horizontal factorization (b) Vertical factorization

Figure 2: Factorizations with message flow

where $\mathscr{P}(Z(1{:}M,1{:}K))$ is the set of all possible state pairs in $Z(1{:}M,1{:}K)$ as given by the graph, and $T(z_i,z_j)$ is the dependency factor capturing the statistical dependency between a pair of nodes. It is an undirected analog to the transition matrix entries in the 2DHMM and can be identified as a two-node potential function in the context of MRFs. Similar to [12], the dependency factors are defined as

$$\begin{aligned}T(m{-}1,m;k) &= \frac{\Pr(Z(m{-}1,k),Z(m,k))}{\Pr(Z(m{-}1,k))\Pr(Z(m,k))},\\ T(k{-}1,k;m) &= \frac{\Pr(Z(m,k{-}1),Z(m,k))}{\Pr(Z(m,k{-}1))\Pr(Z(m,k))}, \qquad (7)\end{aligned}$$

where $T(m{-}1,m;k) := T(Z(m{-}1,k),Z(m,k))$ is used for brevity. This still allows the interpretation as a (scaled) probability. Furthermore, the factor is a symmetric function with respect to the state variables $Z(m,k)$ and does therefore not introduce a directivity as conditional probabilities would do.

Unfortunately, the sum-product algorithm given in [10] is not directly applicable to perform inference, because the given latent state variable structure as shown by Fig. 1 exhibits a grid with tight loops. If only vertical or horizontal connections existed, the model would reduce to several independent chains and the inference problem could be solved by running a modified version of the *forward-backward* algorithm [13] for each chain. However, this assumption would drop the information carried in the neglected direction and is therefore discarded.

Instead, similar approximations as introduced in [7] are used: A local horizontal (i.e. time-directed) tree structure for a fixed frequency index k can be obtained by neglecting all horizontal connections for other frequencies. This results in a single horizontal chain with connected vertical (i.e. frequency-directed) chains as pictured in Fig. 2a. The posterior for a specific node $Z(m,k)$ on the horizontal chain can be factorized into a product of messages under the induced independency assumptions:

$$\begin{aligned}\gamma^{\mathscr{H}}(m,k) \propto\, & p(X(m,k),Z(m,k))\\ & \cdot p(X(1{:}m{-}1,1{:}K)|Z(m,k))\\ & \cdot p(X(m{+}1{:}M,1{:}K)|Z(m,k))\\ & \cdot p(X(m,1{:}k{-}1)|Z(m,k))\\ & \cdot p(X(m,k{+}1{:}K)|Z(m,k)). \qquad (8)\end{aligned}$$

The factors can be interpreted as messages:

$$\gamma^{\mathscr{H}}(m,k) \propto \prod_{d\in\{o,\leftarrow,\rightarrow,\uparrow,\downarrow\}} \mu_d^{\mathscr{H}}(m,k), \qquad (9)$$

ISBN 978-3-8007-4275-2

where the messages are indexed with the message direction or o for an observation message, and the superscript $\mathscr{H}$ denotes the horizontal approximation. Note that the order of messages under the product symbol in Eq. (9) corresponds to the order of terms in Eq. (8) for ease of correspondence. With the exception of the observation message, all messages denote the likelihood of a subgraph of observations, conditioned on the current state variable.

In contrast to [10], the variable-to-factor message and the factor-to-variable message are combined into a single variable-to-variable message type, as each factor node is connected to a maximum of two variable nodes.

If the vertical forward and backward messages are combined with the observation message, the result contains the information of all frequencies for the current time frame:

$$\mu_{\uparrow}^{\mathscr{H}}(m,k)\cdot\mu_{\downarrow}^{\mathscr{H}}(m,k)\cdot\mu_{o}^{\mathscr{H}}(m,k) = p(X(m,1{:}K),Z(m,k)). \tag{10}$$

Using this, ordinary forward and backward runs can be performed along the chain. The forward message can be factorized as follows:

$$\mu_{\rightarrow}^{\mathscr{H}}(m,k) = \sum_{Z(m-1,k)} T(m{-}1,m;k) \prod_{d\in\{\rightarrow,o,\uparrow,\downarrow\}} \mu_{d}^{\mathscr{H}}(m{-}1,k). \tag{11}$$

An equivalent recursion exists for the backward message:

$$\mu_{\leftarrow}^{\mathscr{H}}(m,k) = \sum_{Z(m+1,k)} T(m,m{+}1;k) \prod_{d\in\{\leftarrow,o,\uparrow,\downarrow\}} \mu_{d}^{\mathscr{H}}(m{+}1,k). \tag{12}$$

This scheme can be repeated for every horizontal chain. An equivalent approximation for the vertical direction exists as shown in Fig. 2b with equivalent messages

$$\mu_{d}^{\mathscr{V}}(m,k), \quad d\in\{\uparrow,\downarrow,\leftarrow,\rightarrow,o\}.$$

Using both approximations, each of the messages with $d\in\{\uparrow,\downarrow,\leftarrow,\rightarrow\}$ exists in a 'wide' and a 'narrow' version, e.g.

$$\begin{aligned}\mu_{\rightarrow}^{\mathscr{H}}(m,k) &= p(X(1{:}m{-}1,1{:}K)|Z(m,k)) \quad \text{(wide)},\\ \mu_{\rightarrow}^{\mathscr{V}}(m,k) &= p(X(1{:}m{-}1,k)|Z(m,k)) \quad \text{(narrow)}.\end{aligned}$$

To arrive at an iterative message passing scheme, this distinction is dropped and only the wide versions are kept:

$$\begin{aligned}\mu_{\leftarrow}(m,k) &= \mu_{\leftarrow}^{\mathscr{H}}(m,k), \qquad \mu_{\rightarrow}(m,k) = \mu_{\rightarrow}^{\mathscr{H}}(m,k),\\ \mu_{\uparrow}(m,k) &= \mu_{\uparrow}^{\mathscr{V}}(m,k), \qquad \mu_{\downarrow}(m,k) = \mu_{\downarrow}^{\mathscr{V}}(m,k).\end{aligned}$$

Now, the messages from one direction can be used in turn to provide the replacement observation for the other direction as given by Eq. (10), resulting in an iterative inference scheme. This is an application of the turbo principle known from coding literature [8].

4 Comparison with Directed Model

In [7] a two-dimensional Hidden Markov Model (2DHMM) was proposed, which represents an instance of a Directed Graphical Model, also known as a Bayesian Network. While it was shown to be a viable approach in practice, it suffers from some theoretical caveats:

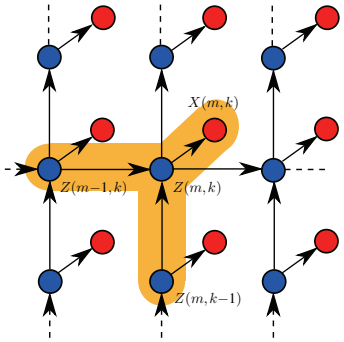


Figure 3: Horizontal factorization of Directed Model showing head-to-head relation

- The Bayesian Network of the Directed Model with the horizontal approximation as shown in Fig. 3 reveals an instance of a head-to-head relation: $Z(m{-}1,k)$ and $Z(m,k{-}1)$ are not statistically independent anymore if $Z(m,k)$ or $X(m,k)$ is observed [11]. Therefore, the factorization proposed in [7], which is equivalent to Eq. (8) above, is not compatible to the graph approximation.
- The directed model implies a semantic of causality in the state sequences because a defined predecessor state exists. While this may be plausible for the time direction, the semantic of a predecessor does not really make sense for the frequency direction.
- With a frequency-dependent transition model, the vertical HMM is inhomogeneous and no stationary distribution exists to be employed as an *a-priori* state probability. Instead, a Markov iteration for every vertical chain has to be executed. Furthermore, the *a-priori* state probabilities in time and frequency domain do not necessarily match, e.g. the transition models are usually contradictory. This property is not carried over to the UGM, where the *a-priori* state probability is defined up-front and the dependency factors depend on it.

Although the UGM is theoretically more appealing, comparison experiments between the two models did not show a significant difference. Nevertheless, we favor the UGM due to its higher flexibility and more compact formulation.

5 Scheduling

The inference scheme described in Section 3 can be viewed as an instance of the sum-product algorithm on a loopy graph with a particular message passing schedule. The message passing is visualized for a small example graph in Fig. 4. Note that only the variable nodes are shown in these timing diagrams. For reasons explained above, this schedule will be referred to as the turbo schedule in the remainder of the paper. While this schedule is justified by the independence assumptions made in Section 3, there are many alternatives one can think of. A popular choice is a flooding schedule mentioned in [10], where every message is updated at every step of the algorithm. Unfortunately, in practice this has shown to be too expensive in terms of runtime for the large graph structure of the SPP task. Instead, two other choices are to be investigated here:

- Horizontal-dominant schedule (see Fig. 5): complete horizontal forward-backward runs with short-chain upward and downward runs in-between. A dual Vertical-

ISBN 978-3-8007-4275-2

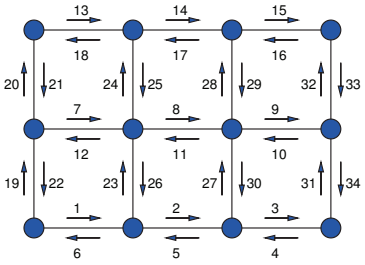


Figure 4: Turbo schedule (numbers indicate order of message passing along the graph)

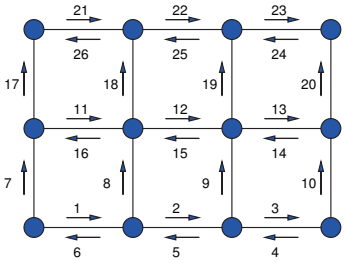


Figure 5: Horizontal dominant schedule (upward only)

dominant schedule is investigated as well.

- Plane-wave schedule (see Fig. 6): The algorithm starts at, e.g., the lower left corner node and computes messages to the right and the upper neighbor node. The next node is the earliest node that receives messages but is not completed. This is done recursively until the upper right corner is reached. Then, the procedure is applied backwards to reach the lower left node.

6 Experiments

The three schedules introduced in the previous Section are to be compared here. Additionally, a decoding assuming independent and identically distributed observations (IID) is performed as a baseline result. The TIMIT database [14] is chosen due to its clean recordings, so that artificially added noise can be freely controlled and ground truth speech is available for comparisons.

For the STFT, a frame size of 1024 and a quarter-frame sized shift are chosen resulting in K=513. A Hamming window is used for analysis. Artificially generated white Gaussian noise is added for the simulation. The amount of additive noise is controlled by the SNR. Binary reference masks are calculated by determining the time-frequency slots of the clean input speech signal that account for 95 % of the cumulative periodogram and setting their mask entry to 1. The dependency factors and the *a-priori* state probabilities are estimated by counting transitions and occurences in the reference masks on all 3260 utterances contributed by male speakers of the TIMIT training corpus while evaluation is done on all 1120 male utterances of the TIMIT test corpus.

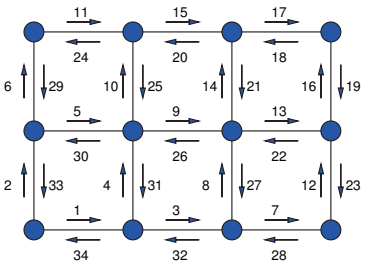


Figure 6: Plane wave schedule

The estimated SPP is compared with the reference mask to obtain the True Positive Rate (TPR) and False Positive Rate (FPR) using the method described in [15], which are utilized to plot a Receiver Operating Characteristic (ROC).

The ROC is shown in Fig. 7 and Fig. 8 for SNRs of −5 dB and 0 dB, respectively. As one can see, all schedule types dominate the IID decision for an FPR exceeding 0.05. Furthermore, the turbo schedule exhibits the best result almost everywhere, with an improvement of approximately 0.08 in terms of the TPR at the knee point for the lower SNR. It is interesting to see that the horizontally-vertically oriented schedules are overall performing better than the diagonally oriented plane-wave schedule.

As a final note, the differences of the vertical-dominant schedule (*VDom*) and the horizontal-dominant schedule (*HDom*) are not significant. In [16], it is suggested to use the direction with the lower variance first — corresponding to the time direction in the SPP task, which is known to exhibit stronger correlations and thus a lower variance. Nonetheless, this does not seem to make a difference for the task at hand.

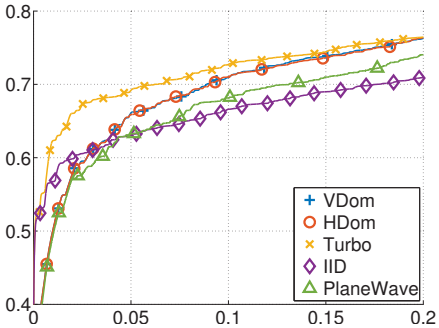


Figure 7: ROC for SNR=−5 dB

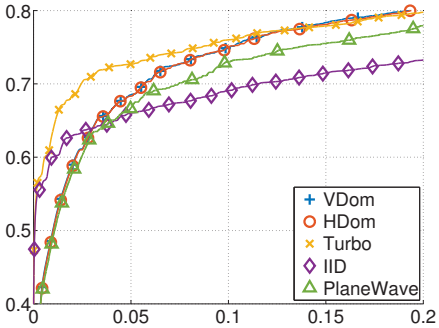


Figure 8: ROC for SNR=0 dB

7 Conclusions

We have presented an Undirected Graphical Model that utilizes information from time and frequency direction in the STFT domain to estimate the Speech Presence Probability. It overcomes the issues of a directed model and allows high flexibility in terms of scheduling. The experimental results compared different schedule schemes and provide strong evidence for the utility of the turbo schedule.

8 Acknowledgement

We thank Prof. P. Höher, CAU Kiel, for fruitful discussions concerning the properties of different scheduling strategies.

References

[1] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4640–4643, May 2011.

[2] J. Benesty, M. M. Sondhi, and Y. A. Huang, *Springer Handbook of Speech Processing*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[3] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, Dec 1984.

[5] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, pp. 910–919, July 2008.

[6] H. Momeni, E. A. P. Habets, and H. R. Abutalebi, "Single-channel speech presence probability estimation using inter-frame and inter-band correlations," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2903–2907, May 2014.

[7] D. H. T. Vu and R. Haeb-Umbach, "Using the turbo principle for exploiting temporal and spectral correlations in speech presence probability estimation," in *38th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pp. 863–867, May 2013.

[8] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes.," in *Communications, 1993. ICC '93 Geneva. Technical Program, Conference Record, IEEE International Conference on*, vol. 2, pp. 1064–1070 vol.2, May 1993.

[9] I. Andrianakis and P. R. White, "On the application of Markov random fields to speech enhancement," in *Proc. IMA Int. Conf. Math. Signal Process*, pp. 198–201, 2006.

[10] F. R. Kschischang, B. J. Frey, and H. A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theor.*, vol. 47, pp. 498–519, Sept. 2006.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[12] S. Receveur, P. Meyer, and T. Fingscheidt, "A compact formulation of turbo audio-visual speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5517–5521, May 2014.

[13] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, Feb 1989.

[14] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus ldc93s1." Web Download, 1993.

[15] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, (New York, NY, USA), pp. 233–240, ACM, 2006.

[16] C. Knievel, Z. Shi, P. A. Hoeher, and G. Auer, "2D graph-based soft channel estimation for MIMO-OFDM," in *Communications (ICC), 2010 IEEE International Conference on*, pp. 1–5, May 2010.

ISBN 978-3-8007-4275-2