USING THE TURBO PRINCIPLE FOR EXPLOITING TEMPORAL AND SPECTRAL CORRELATIONS IN SPEECH PRESENCE PROBABILITY ESTIMATION

Dang Hai Tran Vu and Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany
{tran,haeb}@nt.uni-paderborn.de

ABSTRACT

In this paper we present a speech presence probability (SPP) estimation algorithm which exploits both temporal and spectral correlations of speech. To this end, the SPP estimation is formulated as the posterior probability estimation of the states of a two-dimensional (2D) Hidden Markov Model (HMM). We derive an iterative algorithm to decode the 2D-HMM which is based on the turbo principle. The experimental results show that indeed the SPP estimates improve from iteration to iteration, and further clearly outperform another state-of-the-art SPP estimation algorithm.

1. INTRODUCTION

The estimation of the SPP for each individual time-frequency (TF)-slot in the short-time Fourier transform (STFT) domain is a important part of many speech processing systems. For instance, the widespread speech enhancement approaches based on estimation of the short-time spectral amplitude of the clean speech signal crucially depend on an SPP estimator [1]. However, a reliable SPP estimator is difficult to obtain in a noisy scenario.

It is well known that speech signals have characteristic temporal and spectral correlations in the STFT domain. Usually, this fact is exploited by smoothing the estimated characteristics, such as the SPP estimations themselves, the a priori SNR, or even the gain factor of individual TF bins across time, frequency, or both, e.g., [1–5]. Recently, we proposed a more principled approach to exploiting temporal correlations [6]: Rather than smoothing the estimates with heuristically chosen filter parameters in a postprocessing step, the correlations are directly employed in the estimation of the SPP using a one-dimensional (1D)-HMM for each frequency bin independently.

While this approach was successful in exploiting temporal correlations, an extension to also exploit the spectral correlations asks for the use of 2D-HMMs to capture dependencies over the axes of time and frequency. Unfortunately, exact inference in large 2D-HMMs is computationally infeasible. Even approximative algorithms developed in other fields such as the Markov Chain Monte Carlo approach, incorporate very large computational complexity, e.g., [7].

In the field of telecommunications, the turbo principle has been developed as a powerful means to decode turbo codes [8]. It consists of an iterative decoding process where two conventional decoders exchange soft bit information via an interleaver/deinterleaver pair. In the construction of the turbo decoding scheme much attention is payed to prevent the multiple use of the same information; therefore, the so called *extrinsic* information is identified and exchanged between the two decoders.

Borrowing the ideas of turbo decoding an algorithm is derived here which operates by alternating between horizontal decoding, i.e., decoding along the time axis, and vertical decoding, i.e., along the frequency axis, exploiting temporal correlation in the first and spectral correlation in the second step. *Extrinsic* information is exchanged between the decoders, such that increasingly better estimates of the SPP are obtained. The experimental results show that the proposed decoding algorithm excels with significant performance improvements, high convergence speed and complexity linear in the data set size.

The paper is organized as follows: In Sec. 2 we briefly introduce the observation model used in SPP estimation. Sec. 3 discusses the SPP estimation by considering only correlations in time, thus laying the ground for the extension to the iterative decoding of a 2D-HMM in Sec. 4. Experimental results will then be given in Sec. 5, before we draw conclusion in Sec. 6.

2. SPEECH PRESENCE PROBABILITY ESTIMATION

Let us assume a speech signal S_m^k captured by a microphone as noisy speech signal Y_m^k in the STFT domain, where $m \in \{1, \dots, M\}$ is the time frame index with the utterance length M and $k \in \{1, \dots, K\}$ denotes the frequency bin. With N_m^k denoting the STFT of additive noise we have

$$Y_m^k = \begin{cases} N_m^k & \text{if } Z_m^k = 1 \\ S_m^k + N_m^k & \text{if } Z_m^k = 2 \end{cases}. \quad (1)$$

Here, Z_m^k is binary hidden random variables (RV) indicating whether the TF-slot (m, k) contains noise only ($Z_m^k = 1$) or noisy speech ($Z_m^k = 2$).

It is a common practice in speech enhancement to equivalently model the *a-posteriori* SNR

$$X_m^k = |Y_m^k|^2 / \lambda_m^k, \quad (2)$$

where λ_m^k is the noise variance obtained by a noise tracking algorithm, such as the one in [2]. Assuming zero-mean Gaussian random signals the probability density function (PDF) of the *a-posteriori* SNR X_m^k can be modeled by scaled Chi-squared distributions [3]

$$p\left(X_m^k | Z_m^k = i; \xi_i\right) = \left(\frac{r}{2(1+\xi_i)}\right)^{\frac{r}{2}} \frac{\left(X_m^k\right)^{\frac{r}{2}-1}}{\Gamma\left(\frac{r}{2}\right)} e^{\frac{-X_m^k r}{2(1+\xi_i)}}, \quad (3)$$

where $r = 2$ is the degrees of freedom, $\xi_1 = 0$ for the noise-only case, ξ_2 is the *a-priori* SNR. According to [4] the *a-priori* SNR ξ_2 can be set to a fixed value for a given database.

The goal is to infer the SPP given all *a-posteriori* SNR observations, i.e. to compute the posteriori probability (PP) $\gamma_m^k(i) := P\left(Z_m^k = i | X_{1:M}^{1:K}\right)$, $i = 1, 2$, for all k and m.

3. SPP ESTIMATION USING 1D-HMM ALONG TIME

In this section we briefly review the decoding, i.e., the computation of the PP, of 1D-HMMs. However, we will present the equations in a

This work was supported by DFG under contract number HA 3455/8-1.

slightly different form than is usually done in the literature, e.g., [9]. The reason is to prepare for the extension to the 2D-HMM lattice.

To be able to write the equation in a compact fashion we introduce the following vector and matrix operators: We will write $\mathbf{a} := [\cdot]_{i;j}$ to define the element on the i-th row and j-th column of the matrix $\mathbf{a}$. If $j = 1$ we will omit it. The binary operator $\circ$ is the element-wise product of two vectors, also known as the Hadamard product. Likewise, the binary operator $\oslash$ is the element-wise division and we will write $[\mathbf{a}]^c$ to raise each element of the vector $\mathbf{a}$ to the power of c. The binary rescaling operator of two column vectors of the same size, denoted by the symbol $\propto$, is defined as $\mathbf{a} \propto \mathbf{b} := \mathbf{b}/(\mathbf{a}^{\mathrm{T}}\mathbf{b})$. If the first operand is a scalar then it will be expanded to a column vector of the same size as the second vector operand by repetition. Thus, the operation $1 \propto \mathbf{b}$ rescales the vector $\mathbf{b}$ so that the sum of all elements is one. Throughout this paper we consider the rescaling operator to have the lowest precedence of all operators.

Let us first consider correlations along the time axis only. For each frequency bin k, the sequence of hidden states along the time axis is considered to be a first-order Markov chain. If the state variable Z_m^k is given, the observation X_m^k is independent of all other states and all observations at any other TF slot, see the Bayesian model in Fig. 1(a).

For the sake of simplicity we consider a homogeneous Markov chain, i.e. the 2×2 horizontal transition (HT)-matrix ${}_{\mathcal{H}}\mathbf{T} := [{}_{\mathcal{H}}t(j,i)]_{j;i}$ with entries ${}_{\mathcal{H}}t(j,i) := P\left(Z_m^k = i | Z_{m-1}^k = j\right)$ is independent of m and k. Further we assume that the Markov chains are ergodic and in equilibrium, i.e. the 2×1 column vector of a priori probabilities (APP) $\boldsymbol{\pi} := \left[P\left(Z_m^k = i\right)\right]_{i=1,2}$ is also independent of m and k.

Since we ignore spectral correlations the 2×1 vector of PP can be computed in each frequency bin k independently:

$$\boldsymbol{\gamma}_m^k := \left[P\left(Z_m^k = i | X_{1:M}^k\right)\right]_{i=1,2}. \tag{4}$$

The computation of the PP can be efficiently carried out by employing the forward-backward algorithm (FBA) [10] for each row k separately. In the forward step the 2×1 forward prediction vector (FPV)

$$\boldsymbol{\alpha}_m^k := \left[p\left(X_{1:m-1}^k, Z_m^k = i\right) \Big/ p\left(X_{1:m-1}^k\right)\right]_{i=1,2} \tag{5}$$

and in the backward step the 2×1 backward vector (BV)

$$\boldsymbol{\beta}_m^k := \left[p\left(X_{m+1:M}^k | Z_m^k = i\right) \Big/ p\left(X_{m+1:M}^k\right)\right]_{i=1,2} \tag{6}$$

is computed by combining observation evidence vector (OEV)

$$\mathbf{o}_m^k := \boldsymbol{\pi} \propto \left[p\left(X_m^k | Z_m^k = i\right)\right]_{i=1,2}, \tag{7}$$

with ${}_{\mathcal{H}}\mathbf{T}$ and $\boldsymbol{\pi}$ using dynamic programming techniques. Using the introduced matrix operators the FBA can be compactly written as:

$$\boldsymbol{\alpha}_m^k = 1 \propto {}_{\mathcal{H}}\mathbf{T}^{\mathrm{T}}\left(\boldsymbol{\alpha}_{m-1}^k \circ \mathbf{o}_{m-1}^k\right), \tag{8}$$

$$\boldsymbol{\beta}_m^k = \boldsymbol{\pi} \propto {}_{\mathcal{H}}\mathbf{T}\left(\boldsymbol{\beta}_{m+1}^k \circ \mathbf{o}_{m+1}^k\right), \tag{9}$$

$$\boldsymbol{\gamma}_m^k = 1 \propto \underbrace{\boldsymbol{\pi}}_{\textit{prior}} \circ \underbrace{\mathbf{o}_m^k}_{\textit{intrinsic}} \circ \underbrace{\left(\boldsymbol{\alpha}_m^k \oslash \boldsymbol{\pi}\right) \circ \boldsymbol{\beta}_m^k}_{\textit{extrinsic}}, \tag{10}$$

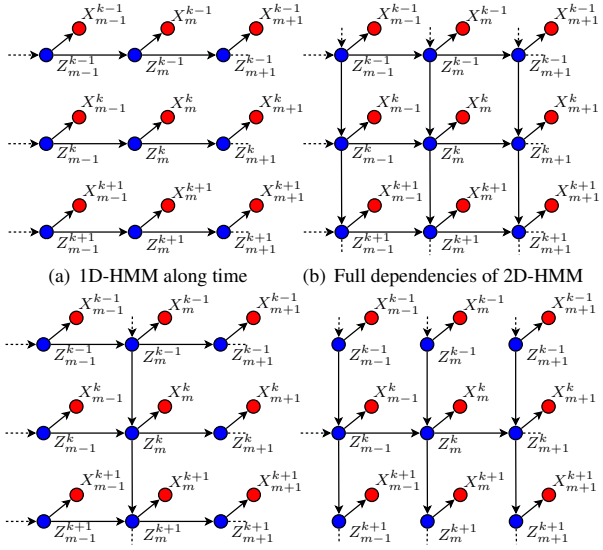
where $\boldsymbol{\alpha}_1^k = \boldsymbol{\pi}$ and $\boldsymbol{\beta}_M^k = [1,1]^{\mathrm{T}}\ \forall k$.

We see from equation (10) that the PP is a product of three terms. The first term is the state *prior* distribution. The second term might be called *intrinsic* information since it contains the evidence of an individual observation in the given TF-slot. The third term, is the *extrinsic* information which takes into account the knowledge gained from past observations, concentrated in $\boldsymbol{\alpha}_m^k \oslash \boldsymbol{\pi}$, and the future observations, concentrated in $\boldsymbol{\beta}_m^k$. The distinction of these three terms is key to understanding of the turbo decoding scheme.

Beside numerical stability the benefit of the used normalization of $\mathbf{o}_m^k$, $(\boldsymbol{\alpha}_m^k \oslash \boldsymbol{\pi})$ and $\boldsymbol{\beta}_m^k$ is an easier interpretation of these values in terms of information gain. If the current observation contains no information about the state Z_m^k then the OEV is $\mathbf{o}_m^k = [1,1]^{\mathrm{T}}$ and if no information can be yielded from past or future observations then the FPV and BV terms are $\boldsymbol{\alpha}_m^k \oslash \boldsymbol{\pi} = [1,1]^{\mathrm{T}}$ or $\boldsymbol{\beta}_m^k = [1,1]^{\mathrm{T}}$, respectively. The logarithm to base 2 of these quantities is also known as the pointwise mutual information. Following this link to information theory the denotation of the three terms can be further justified by taking the expectation of the negative logarithm to the base 2 of equation (10) on both sides w.r.t. Z_m^k and $X_{1:M}^k$. Recalling the definition of entropy $H(u) := \mathrm{E}\left[-\log_2(p(u))\right]$ and mutual information $I(u;v) := \mathrm{E}\left[\log_2(p(u,v)/(p(u)\,p(v)))\right]$ we have

$$H(Z_m^k | X_{1:M}^k) = \overbrace{H(Z_m^k)}^{\textit{prior}} - \overbrace{I(Z_m^k; X_m^k)}^{\textit{intrinsic}} - \underbrace{I(Z_m^k; X_{1:m-1}^k) - I(Z_m^k; X_{m+1:M}^k)}_{\textit{extrinsic}} + R(X_{1:M}^k). \tag{11}$$

Thus, the uncertainty about hidden state Z_m^k after observing $X_{1:M}^k$ is equal to the uncertainty before the observation minus the information of the current observation and the information obtained from past and future observations. The last term $R(X_{1:M}^k)$ in equation (11) bears the redundant information in the observations.



(a) 1D-HMM along time (b) Full dependencies of 2D-HMM

(c) Vertical pass in m-th column (d) Horizontal pass in k-th row

Fig. 1. *Bayesian model depicting statistical dependencies.*

4. ITERATIVE DECODING OF 2D HMM

Now we make the extension from 1D-HMMs to a 2D-HMM by regarding temporal and spectral correlations. For convenience we write $Z_m^{1:K}$ to denote the m-th column and $Z_{1:M}^k$ to denote the k-th row of the hidden RV, respectively. Likewise, we use the same notation for the lattice of observation vectors $X_{1:M}^{1:K}$. Now we consider Z_m^k as a 2D random Markov process as is depicted in Fig. 1(b). Again, a homogeneous and ergodic Markov process in equilibrium is assumed. Thus, the APP vector is $\boldsymbol{\pi} := \left[P\left(Z_m^k = i\right)\right]_{i=1,2}\ \forall m,k$. The 2D-HMM requires the specification of a 3D transition matrix with ${}_{3\mathrm{D}}t(j_1,j_2,i) := P(Z_m^k = i | Z_{m-1}^k = j_1, Z_m^{k-1} = j_2)$. Similar to [11] and [12] we reduce the complexity of the model by assuming that this transition matrix is *separable*, i.e. it can be decomposed

into a product of horizontal transitions ${}_{\mathcal{H}}t(j,i)$ and vertical transitions ${}_{\mathcal{V}}t(j,i) := P\left(Z_m^k = i | Z_m^{k-1} = j\right)$. Hence, we have

$${}_{3\mathrm{D}}t(j_1, j_2, i) = {}_{\mathcal{H}}t(j_1, i)\,{}_{\mathcal{V}}t(j_2, i) / \sum_{\tilde{i}=1}^{S} {}_{\mathcal{H}}t(j_1, \tilde{i})\,{}_{\mathcal{V}}t(j_2, \tilde{i}). \quad (12)$$

The vertical transition probabilities are collected in a vertical transition (VT)-matrix ${}_{\mathcal{V}}\mathbf{T}$ with elements $\left[{}_{\mathcal{V}}t(j,i)\right]_{j;i}$, similar to ${}_{\mathcal{H}}\mathbf{T}$. Note, that while temporal correlations are stored in the HT-matrix, spectral correlations are stored in VT-matrix.

Decoding a large 2D-HMM, i.e., computing the PP vector $\boldsymbol{\gamma}_m^k := \left[P\left(Z_m^k = i | X_{1:M}^{1:K}\right)\right]_{i=1,2}$ is computationally infeasible and no efficient algorithms are known for an exact solution. The reason for the difficulties compared to an 1D-HMM is that no single state exists that *d-separates* the graphical model into independent sets of vertices , i.e. the Bayesian model in Fig. 1(b) is not a poly-tree [13].

To derive an approximate algorithm we propose to split the decoding into horizontal and vertical processing steps and let the steps exchange information by inducing additional information on each other.

Let us derive the vertical processing (VP)-step where we decode the 2D-HMM column-by-column, but also account for information in the rows. For the m-th column we ignore the vertical dependencies in all other columns. Therefore, the VP-steps are independent from each other. Fig. 1(c) depicts the statistical dependencies of VP-step in m-th column. Reflecting the notation of the 1D-HMM we first introduce the following 2×1 column vectors

$${}_{\mathcal{V}}\boldsymbol{\gamma}_m^k := \left[P\left(Z_m^k = i | X_{1:M}^{1:K}\right)\right]_{i=1,2}, \quad (13)$$

$${}_{\mathcal{V}}\boldsymbol{\alpha}_m^k := \left[p\left(X_{1:M}^{1:k-1}, Z_m^k = i\right) \Big/ p\left(X_{1:M}^{1:k-1}\right)\right]_{i=1,2}, \quad (14)$$

$${}_{\mathcal{V}}\boldsymbol{\beta}_m^k := \left[p\left(X_{1:M}^{k+1:K} | Z_m^k = i\right) \Big/ p\left(X_{1:M}^{k+1:K}\right)\right]_{i=1,2}, \quad (15)$$

$${}_{\mathcal{V}}\mathbf{u}_m^k := \left[\frac{p\left(X_{1:m-1}^k | Z_m^k = i\right) p\left(X_{m+1:M}^k | Z_m^k = i\right)}{p\left(X_{1:m-1}^k\right) p\left(X_{m+1:M}^k\right)}\right]_{i=1,2}, \quad (16)$$

where ${}_{\mathcal{V}}\boldsymbol{\gamma}_m^k$ is the vertical PP, ${}_{\mathcal{V}}\boldsymbol{\alpha}_m^k$ is the vertical FPV, ${}_{\mathcal{V}}\boldsymbol{\beta}_m^k$ is the vertical BV and ${}_{\mathcal{V}}\mathbf{u}_m^k$ is the vertical junction vector (JV). Now, given the simplified model in Fig. 1(c) all vertices Z_m^k of an active column *d-separate* the graphical model into five sets corresponding to states and observations above the current state $\{Z_{1:M}^{1:k-1}, X_{1:M}^{1:k-1}\}$, below the current state $\{Z_{1:M}^{k+1:K}, X_{1:M}^{k+1:K}\}$, left of the current state $\{Z_{1:m-1}^k, X_{1:m-1}^k\}$, right of the current state $\{Z_{m+1:M}^k, X_{m+1:M}^k\}$ and finally $\{X_m^k\}$. Hence, the joint PDF can factorized as

$$p\left(Z_m^k, X_{1:M}^{1:K}\right) = P\left(Z_m^k\right) p\left(X_{1:M}^{1:k-1} | Z_m^k\right) p\left(X_{1:M}^{k+1:K} | Z_m^k\right) \cdot p\left(X_{1:m-1}^k | Z_m^k\right) p\left(X_{m+1:M}^k | Z_m^k\right) p\left(X_m^k | Z_m^k\right). \quad (17)$$

By using this property the vertical FPV and vertical BV can be recursively computed by a slightly modified version of the FBA:

$${}_{\mathcal{V}}\boldsymbol{\alpha}_m^k = 1 \propto {}_{\mathcal{V}}\mathbf{T}^{\mathrm{T}} \left({}_{\mathcal{V}}\boldsymbol{\alpha}_m^{k-1} \circ \mathbf{o}_m^{k-1} \circ {}_{\mathcal{V}}\mathbf{u}_m^{k-1}\right), \quad (18)$$

$${}_{\mathcal{V}}\boldsymbol{\beta}_m^k = \boldsymbol{\pi} \propto {}_{\mathcal{V}}\mathbf{T} \left({}_{\mathcal{V}}\boldsymbol{\beta}_m^{k+1} \circ \mathbf{o}_m^{k+1} \circ {}_{\mathcal{V}}\mathbf{u}_m^{k+1}\right), \quad (19)$$

where ${}_{\mathcal{V}}\boldsymbol{\alpha}_m^1 = \boldsymbol{\pi}$ and ${}_{\mathcal{V}}\boldsymbol{\beta}_m^K = [1,1]^{\mathrm{T}}$.

Note, that the JV ${}_{\mathcal{V}}\mathbf{u}_m^k$ is equal to the extrinsic factor in equation (10) of the k-th 1D-HMM in horizontal direction:

$${}_{\mathcal{V}}\mathbf{u}_m^k = (\boldsymbol{\alpha}_m^k \oslash \boldsymbol{\pi}) \circ \boldsymbol{\beta}_m^k. \quad (20)$$

As already implied in (11) the JV consequently keeps track of all information from past and future observations of each individual frequency row. It is easy to verify that if there is no information in the temporal chains corresponding to ${}_{\mathcal{V}}\mathbf{u}_m^k = [1,1]^{\mathrm{T}}$ then the modified FBA in (18) and (19) is equal to the ordinary FBA along the spectral dependencies.

After applying the modified FBA the PP of the m-th row ${}_{\mathcal{H}}\boldsymbol{\gamma}_m^k$ can be obtained by

$${}_{\mathcal{V}}\boldsymbol{\gamma}_m^k = 1 \propto \underbrace{\boldsymbol{\pi}}_{\textit{prior}} \circ \underbrace{\mathbf{o}_m^k \circ {}_{\mathcal{V}}\mathbf{u}_m^k}_{\textit{intrinsic}} \circ \underbrace{({}_{\mathcal{V}}\boldsymbol{\alpha}_m^k \oslash \boldsymbol{\pi}) \circ {}_{\mathcal{V}}\boldsymbol{\beta}_m^k}_{\textit{extrinsic}}, \quad (21)$$

which is already suggested by the factorization in (17).

The JV ${}_{\mathcal{V}}\mathbf{u}_m^k$ can be viewed as an additional independent observation evidence vector. Thus, it must be considered as an *intrinsic* factor in the PP formula (21). The *prior* and the *extrinsic* terms are just as in equation (10).

Since the 2D-HMM is symmetric a similar set of equations can be derived for a horizontal processing (HP) ignoring all other horizontal dependencies except for the considered row by substituting the indices in formulas of VP, see Fig. 1(d). Analog to (13) - (16) we define the horizontal FPV ${}_{\mathcal{H}}\boldsymbol{\alpha}_m^k$, the horizontal BV ${}_{\mathcal{H}}\boldsymbol{\beta}_m^k$, the horizontal JV ${}_{\mathcal{H}}\mathbf{u}_m^k$ and the horizontal PP ${}_{\mathcal{H}}\boldsymbol{\gamma}_m^k$. The modified FBA for the horizontal processing is given by:

$${}_{\mathcal{H}}\boldsymbol{\alpha}_m^k = 1 \propto {}_{\mathcal{H}}\mathbf{T}^{\mathrm{T}} \left({}_{\mathcal{H}}\boldsymbol{\alpha}_{m-1}^k \circ \mathbf{o}_{m-1}^k \circ {}_{\mathcal{H}}\mathbf{u}_{m-1}^k\right), \quad (22)$$

$${}_{\mathcal{H}}\boldsymbol{\beta}_m^k = \boldsymbol{\pi} \propto {}_{\mathcal{H}}\mathbf{T} \left({}_{\mathcal{H}}\boldsymbol{\beta}_{m+1}^k \circ \mathbf{o}_{m+1}^k \circ {}_{\mathcal{H}}\mathbf{u}_{m+1}^k\right), \quad (23)$$

$${}_{\mathcal{H}}\boldsymbol{\gamma}_m^k = 1 \propto \underbrace{\boldsymbol{\pi}}_{\textit{prior}} \circ \underbrace{\mathbf{o}_m^k \circ {}_{\mathcal{H}}\mathbf{u}_m^k}_{\textit{intrinsic}} \circ \underbrace{({}_{\mathcal{H}}\boldsymbol{\alpha}_m^k \oslash \boldsymbol{\pi}) \circ {}_{\mathcal{H}}\boldsymbol{\beta}_m^k}_{\textit{extrinsic}}, \quad (24)$$

where ${}_{\mathcal{H}}\boldsymbol{\alpha}_1^k = \boldsymbol{\pi}$ and ${}_{\mathcal{H}}\boldsymbol{\beta}_M^k = [1,1]^{\mathrm{T}}$.

The key to improve the modified FBA are the JVs ${}_{\mathcal{V}}\mathbf{u}_m^k$ and ${}_{\mathcal{H}}\mathbf{u}_m^k$. Due to the approximation in the simplified model in Fig. 1(c) and 1(d) they are computed by ignoring the other horizontal or vertical dependencies; therefore, it is apparent that they are suboptimal. The core of the proposed algorithm is to use the *extrinsic* term of the previous VP-step as the JV ${}_{\mathcal{V}}\mathbf{u}_m^k$ in the HP-step:

$${}_{\mathcal{H}}\mathbf{u}_m^k \leftarrow ({}_{\mathcal{V}}\boldsymbol{\alpha}_m^k \oslash \boldsymbol{\pi}) \circ {}_{\mathcal{V}}\boldsymbol{\beta}_m^k. \quad (25)$$

Subsequently, the *extrinsic* factor of the HP-step is used as the augmented JV ${}_{\mathcal{H}}\mathbf{u}_m^k$ to rerun the VP-step:

$${}_{\mathcal{V}}\mathbf{u}_m^k \leftarrow ({}_{\mathcal{H}}\boldsymbol{\alpha}_m^k \oslash \boldsymbol{\pi}) \circ {}_{\mathcal{H}}\boldsymbol{\beta}_m^k. \quad (26)$$

After that we start over again and thus arrive at an iterative decoding scheme.

The reason for forwarding the *extrinsic* factor, rather than the PP, to the next processing step is to prevent the reuse of information in the next iteration. Suppose that we forwarded the PP instead of the *extrinsic* factor, as proposed in [12]. By plugging the equations (24) and (21) into each other it is easy to verify that the hypothetical final PP ${}_{\star}\boldsymbol{\gamma}_m^k$ has the form

$${}_{\star}\boldsymbol{\gamma}_m^k = 1 \propto \left([\boldsymbol{\pi}]^{2C} \circ [\mathbf{o}_m^k]^{2C} \circ \cdots\right), \quad (27)$$

where we omit the complicated *extrinsic* terms and where C is the number of full vertical and horizontal processing cycles. From (27) it is apparent that forwarding the PP into the next iteration may lead to faster convergence, but also to biased results because the *prior* and the *intrinsic* terms are regarded multiple times.

The prevention of information reuse is also the reason for the requirement that the JVs ${}_{\mathcal{V}}\mathbf{u}_m^k$ and ${}_{\mathcal{H}}\mathbf{u}_m^k$ should be statistically independent of the observation evidence. For our application, however, forwarding the *extrinsic* factor merely attenuates the reuse of information since, compared to the turbo coding, we have no interleaver between the processing steps. Thus, independence of the observation evidence can only be guaranteed in the first processing cycle.

The derived iterative decoding algorithm can be reformulated to be a special case of the sum-product-algorithm, [14, 15], with a "right-left-down-up" message passing schedule. To this end it is hard to prove that the proposed iterative turbo decoding scheme converges to the optimal or stable solution, i.e. that the horizontal and vertical PPs agree with each other or that the induced JVs ${}_{\mathcal{V}}\mathbf{u}_m^k$ and

${}_{\mathcal{H}}\mathbf{u}_m^k$ reach stable states between VP and HP-steps. But we can provide at least a necessary criterion for the PPs of the two precessing directions to coincide: The VT and HT matrices must have the same principal eigenvector $\boldsymbol{\pi}$

$$\boldsymbol{\pi}\ {}_{\mathcal{V}}\lambda = {}_{\mathcal{V}}\mathbf{T}^{\mathrm{T}}\ \boldsymbol{\pi} \qquad \text{and} \qquad \boldsymbol{\pi}\ {}_{\mathcal{H}}\lambda = {}_{\mathcal{H}}\mathbf{T}^{\mathrm{T}}\ \boldsymbol{\pi}, \tag{28}$$

where ${}_{\mathcal{V}}\lambda$ and ${}_{\mathcal{H}}\lambda$ are the corresponding eigenvalues. All other eigenvalues of the positive-definite matrices ${}_{\mathcal{H}}\mathbf{T}^{\mathrm{T}}$ and ${}_{\mathcal{V}}\mathbf{T}^{\mathrm{T}}$ must be smaller than ${}_{\mathcal{V}}\lambda$ and ${}_{\mathcal{H}}\lambda$, respectively. This condition is essential for the proposed iterative algorithm since otherwise the 2D-HMM is not ergodic and has no equilibrium.

The need for this criterion to hold can be easily verified by the following thought experiment. Suppose that ${}_{\mathcal{V}}\mathbf{T}^{\mathrm{T}}$ and ${}_{\mathcal{H}}\mathbf{T}^{\mathrm{T}}$ have two different principal eigenvectors ${}_{\mathcal{V}}\boldsymbol{\pi}$ and ${}_{\mathcal{H}}\boldsymbol{\pi}$, respectively. Furthermore let all observations contain no information which is equivalent to $\mathbf{o}_m^k = [1,1]^{\mathrm{T}} \forall m,k$. Then the PP of the VP is ${}_{\mathcal{V}}\boldsymbol{\pi}$ and the PP of the HP is ${}_{\mathcal{H}}\boldsymbol{\pi}$. Thus, there are no unambiguous solutions.

To define a stopping condition for the iterative decoding one could measure the changes of the PP after a full HP and VP cycle using the Kullback-Leibler divergence and stop iterating when no significant changes have occurred. However, in practice we observe that after three or four cycles the PP is in steady state. Therefore, we recommend a fixed number of cycles C for SPP estimation.

Taken all together the proposed method is summarized in the Algorithm 1 box.

Algorithm 1 Iterative turbo decoding of 2D-HMM

- Set ${}_{\mathcal{H}}\mathbf{u}_m^k \leftarrow [1,1]^{\mathrm{T}} \forall m,k$.

for $c = 1$ to C **do**

- HP-step given in equation (22) and (23).
- Set ${}_{\mathcal{V}}\mathbf{u}_m^k \leftarrow ({}_{\mathcal{H}}\boldsymbol{\alpha}_m^k \oslash \boldsymbol{\pi}) \circ {}_{\mathcal{H}}\boldsymbol{\beta}_m^k$.
- VP-step given in equation (18) and (19).
- Set ${}_{\mathcal{H}}\mathbf{u}_m^k \leftarrow ({}_{\mathcal{V}}\boldsymbol{\alpha}_m^k \oslash \boldsymbol{\pi}) \circ {}_{\mathcal{V}}\boldsymbol{\beta}_m^k$.

end for

- Set final SPP to $\boldsymbol{\gamma}_m^k \leftarrow {}_{\mathcal{V}}\boldsymbol{\gamma}_m^k$ using equation (21).

5. EXPERIMENTS

For the experiments we considered 40 clean speech utterances (20 male, 20 female) taken from the TIMIT database sampled at 16 kHz. Four different noise types (white, pink, babble, volvo) taken from the NOISEX-92 database were added to the speech signals at five different SNR levels (0dB, 5dB, 10dB) to obtain the noisy speech signals. To remove anomalies around 0 Hz, especially in the 'volvo'-noise file, we preprocessed all signals with an 80 Hz - 7.8 kHz IIR band pass filter. The signals are converted into the STFT-domain using a 1024-point Hann window with 512 samples frame shift. We used our own implementation of the minimum statistics based noise tracker to estimate the noise variance λ_m^k.

To evaluate the performance the following procedure is applied: The estimated SPPs are quantized towards a speech presence decision mask $\tilde{\gamma}_m^k$ using a threshold δ where $\delta = 0.5$ is equivalent to a maximum *a-posteriori* (MAP) decision:

$$\tilde{\gamma}_m^k = \begin{cases} 1 & \text{if } P\left(Z_m^k = 2 | X_{1:M}^{1:K}\right) > \delta \\ 0 & \text{else.} \end{cases} \tag{29}$$

Then, the detection rate (DR) and the false alarm rate (FAR), averaged over all TF-slots and all noise types and SNR conditions, are computed. The ground-truth speech presence reference masks for DR/FAR computation are created by marking all TF-slots as "speech present", where its energy belongs to the 99.9 percent quantile of the total energy of a given utterance in each frequency bin. To obtain a receiver operating characteristic (ROC) curve the decision threshold δ is sweeped through the interval $[0,1]$. ROC curves of the proposed algorithm for different HP and VP iteration cycles C are computed, where the 0th cycle corresponds to a SPP detector ignoring all dependencies, i.e. the states Z_m^k are assumed to be independent and identically distributed (i.i.d.). We used the SPP estimator according to Gerkmann et al. [4] with non-causal local smoothing window for reference. In Fig. 2 the ROC curves are depicted.

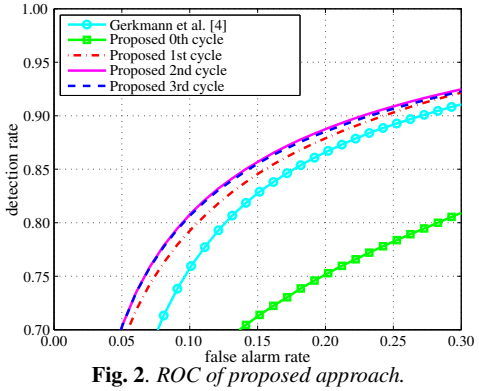


Fig. 2. *ROC of proposed approach.*

From Fig. 2 it can be noticed that the performance gain from the 0th cycle to the 1st cycle is huge. Compared to the local smoothing in [4] our algorithm exhibits a significantly better performance even after just one cyle. We observe moderate gains in the 2nd cycle and no significant changes afterwards. We can thus conclude that the algorithm converges quickly.

In our experiments we found out that the choice of the parameter ξ_2 in the observation model (3) is critical and must be tuned by hand towards relatively low values ($\xi_2 = 4$dB). We conjecture that a high value of the *a-priori* SNR leads to high confidence in the observation evidences. Consequently, the contributions of the neighboring states are too weak, and the modified FBA will never revise its decision and favor an alternative hypothesis.

6. CONCLUSIONS

In this paper we employed a 2D-HMM to capture temporal and spectral correlations in the time-frequency domain for an improved speech presence probability estimation. We derived a modified FBA for iterative decoding of 2D-HMMs, which is based on the turbo principle. The resulting inference algorithm iteratively alternates between vertical and horizontal processing steps. A key contribution is to define and isolate the extrinsic information to be exchanged between horizontal and vertical decoding. We confirmed by experiments that the decoding scheme results in an improved performance compared to a widely used SPP estimator.

Future work will be devoted to a fixed-lag implementation to arrive at an SPP decision for every incoming STFT frame after only a low latency.

7. RELATION TO PRIOR WORK

We presented an novel approach for speech presence probability estimation exploiting correlations of adjacent time-frequency slots. Usually, these correlations are exploited by some heuristic smoothing techniques, e.g. [1–5]. The approach proposed here is more rigorous. Based on our previous work with one dimensional HMMs along the time axis, [6], we extend the modeling to 2D-HMMs to exploit both temporal and spectral correlations in the speech signal.

8. REFERENCES

[1] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP1999)*, Mar 1999, vol. 2, pp. 789 –792.

[2] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504 –512, Jul 2001.

[3] R. Martin and T. Lotter, "Optimal recursive smoothing of non-stationary periodograms," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2001.

[4] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 910–919, Jul. 2008.

[5] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 5, pp. 2403–2418, 2001.

[6] D. H. Tran Vu and R. H. Haeb-Umbach, "Exploiting temporal correlations in joint multichannel speech separation and noise suppression using hidden Markov models," in *International Workshop on Acoustic Signal Enhancement (IWAENC2012)*, Sep. 2012.

[7] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721 –741, Nov. 1984.

[8] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes (1)," in *IEEE International Conference on Communications(ICC1993)*, May 1993, vol. 2, pp. 1064 –1070.

[9] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[10] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate (corresp.)," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 284 – 287, Mar 1974.

[11] H. Othman and T. Aboulnasr, "A separable low complexity 2D HMM with application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1229 – 1238, Oct. 2003.

[12] F. Perronnin, J.-L. Dugelay, and K. Rose, "Iterative decoding of two-dimensional hidden Markov models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2003)*, April 2003, vol. 3, pp. 329–332.

[13] D. Geiger, T. Vera, and J. Pearl, "Identifiying independence in Bayesian networks," *Networks*, vol. 20, pp. 507–534, 1990.

[14] S. M. Aji and R. J. McEliece, "The generalized distributive law," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 325 –343, Mar 2000.

[15] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498 –519, Feb 2001.