Improved Speech Presence Probabilities Using HMM-Based Inference, with Applications to Speech Enhancement and ASR

Bengt J. Borgström, *Student Member, IEEE,* and Abeer Alwan, *IEEE Fellow*

***Abstract*— This paper presents a technique for determining improved speech presence probabilities (SPPs), by exploiting the temporal correlation present in spectral speech data. Based on a set of traditional SPPs, we estimate the underlying speech presence probability via statistical inference. Traditional SPPs are assumed to be observations of channel-specific two-state Markov models. Corresponding steady-state and transitional statistics are set to capture the well-known temporal correlation of spectral speech data, and observation statistics are modeled based on the effect of additive acoustic noise on resulting SPPs. Once underlying models have been parameterized, improved speech presence probabilities can be estimated via traditional inference techniques, such as the forward or forward-backward algorithms. The 2-state configuration of underlying signal models enables low complexity HMM-based processing, only slightly increasing complexity relative to standard SPPs, and thereby making the proposed framework attractive for resource-constrained scenarios. Proposed SPP masks are shown to provide a significant increase in accuracy relative to the state-of-the-art method of [12], in terms of the mean pointwise Kullback-Leibler (KL) distance. When applied to soft-decision speech enhancement, proposed SPPs show improved results in terms of segmental SNRs. Closer analysis reveals significantly decreased noise leakage, whereas speech distortion is increased. When applied to automatic speech recognition (ASR), the use of soft-decision enhancement with proposed SPPs provides increased recognition performance, relative to [12].**

***Index Terms*— Speech Presence Probability, Noise Suppression, Soft-Decision Speech Enhancement, Automatic Speech Recognition, Hidden Markov Models.**

I. Introduction

Speech communication in adverse acoustic environments is an inherently difficult problem due in part to the corruptive effect of background noise. Whether for human-to-human interaction, or for human to machine platforms, single channel noise suppression of speech signals offers a computationally efficient option for improving the effectiveness of communication in such environments. The issue of complexity is especially important when designing mobile devices, which tend to be resource-constrained. This paper addresses the problem of single-channel noise suppression and proposes a low complexity algorithm which is shown to be effective for speech enhancement and ASR.

Spectral masks serve as pivotal tools in noise robust speech processing systems, since they provide valuable information regarding speech presence uncertainty throughout spectro-temporal locations. Speech presence probability (SPP) values have been widely used to derive minimum mean-square (MMSE) soft-decision spectral speech enhancement algorithms ([1]-[3]), which can be utilized to improve perceptual quality or noise robust automatic speech recognition (ASR) rates. Additionally, statistical model-based voice activity detection (VAD) algorithms typically utilize a theory related to SPPs when determining active speech regions ([6],[17],[18]).

The calculation of traditional speech presence probabilities builds upon statistical modeling of speech and noise spectral components [19]. The framework for determining SPPs is robust to the exact models used for speech and noise, and various studies have explored the use of Gaussian distributions ([1]-[3]), Laplacian distributions ([16],[20]), and others. A drawback of traditional SPP masks is that they base probabilities strictly on time-specific observations, and fail to exploit the well-known temporal correlation of time-frequency speech data.

Improvements to traditional SPPs for the task of speech enhancement have been presented in [12] and [22]. In [12], the authors explicitly determine separate speech presence probabilities on local, global, and frame levels, and combine them in a soft-decision manner. Similarly, in [22], the authors integrate local and global information, after smoothing observed spectra along the temporal and frequency axes.

In this paper, we propose a framework for determining improved SPPs using statistical inference. We provide motivation for the assumption of standard SPPs as observations of channel-specific 2-state Markov models, wherein the states represent active and inactive spectral speech regions, respectively. Corresponding steady-state transitional statistics are set to capture the well-known temporal correlation of spectral speech data, and observation statistics can be modeled based on the effect of additive acoustic noise on resulting SPP values. Once underlying models have been parameterized, improved speech presence probabilities can be estimated via traditional HMM-based decoding techniques, such as the forward and forward-backward algorithms [4]. The proposed method is shown to significantly decrease the mean pointwise KL distance between masks obtained from noisy speech, and those obtained from corresponding "oracle" signals, relative to methods in [2] and [12].

A well-known downside to HMM-based inference is the induced computational load. This is especially problematic for speech processing applications which may be resource-constrained and/or delay-sensitive. However, the proposed method for determining improved SPPs induces a small computational load relative to standard SPPs, due to the underlying

model. That is, since the underlying channel-specific signal models are comprised of only 2 states, HMM-based decoding is quite efficient.

Traditional statistical noise suppression rules for speech operate under the assumption of speech presence throughout spectro-temporal locations. However, this assumption is not valid both for time periods of inactive speech, and for frequency channels corresponding to harmonic valleys or vocal tract zeros. Instead, utilizing speech presence probability during noise suppression takes into account speech presence uncertainty, and leads to minimum mean-square error (MMSE) soft-decision speech enhancement ([1]-[3],[12],[14],[20]). The proposed decoding technique is applied to soft-decision speech enhancement to illustrate its effectiveness. HMM-based decoding is shown to improve soft-decision speech enhancement relative to the state-of-the-art method in [12], in terms of segmental SNR (SSNR). Closer analysis reveals reduced noise leakage while maintaining low speech distortion. Additionally, proposed masks provide increased performance for automatic speech recognition (ASR). Complexity analysis reveals that proposed framework in this paper to be less complex than those in [12].

This paper is organized as follows: the standard framework for determining speech presence probabilities, which utilizes channel-specific generalized likelihood ratios (GLRs), is reviewed in Section II. Improved SPPs using HMM-based inference are presented in Section III. Section IV provides experimental results for SPP mask accuracy as well as soft-decision speech enhancement and ASR. Conclusions are presented in Section V.

II. Speech Presence Probabilities

In this section, we review the standard framework for determining speech presence probabilities, which utilizes channel-specific generalized likelihood ratios (GLRs). Throughout this study, we assume an additive noisy speech model. Using a stochastic approach, this can be expressed in the spectral domain as:

$$E\left[|x_m(n)|^2\right] = E\left[|s_m(n)|^2\right] + E\left[|d_m(n)|^2\right], \quad (1)$$

where $x_m(n)$ is the observed spectral data, $d_m(n)$ is the corruptive noise, $s_m(n)$ is the underlying clean speech, m denotes frequency channel index, and n denotes time index. During SPP mask estimation, we have access to the observed spectral data as well as an estimate of the local noise variance, $\lambda_{d,m}(n)$. We define the *a priori* and *a posteriori* SNRs, $\xi_m(n)$ and $\gamma_m(n)$ respectively, as in [1]:

$$\xi_m(n) = \frac{E\left[|s_m(n)|^2\right]}{E\left[|d_m(n)|^2\right]} = \frac{\lambda_{s,m}(n)}{\lambda_{d,m}(n)} \quad (2)$$

$$\gamma_m(n) = \frac{|x_m(n)|^2}{E\left[|d_m(n)|^2\right]} = \frac{|x_m(n)|^2}{\lambda_{d,m}(n)}. \quad (3)$$

Furthermore, the decision-directed (DD) approach is used to approximate $\xi_m(n)$ [2].

In deriving SPP masks, we assume that spectral masks are observations of channel-specific 2-state information sources. That is, for a given time index, the spectral coefficient corresponding to frequency channel m is either produced during an inactive speech region, denoted by H_m^0, or by an active speech region, denoted by H_m^1:

$$\begin{cases} H_m^0(n): & E\left[|x_m(n)|^2\right] = E\left[|d_m(n)|^2\right] \\ H_m^1(n): & E\left[|x_m(n)|^2\right] = E\left[|s_m(n)|^2\right] + E\left[|d_m(n)|^2\right] \end{cases}, \quad (4)$$

$$\text{for } 1 \le m \le N_m,$$

where N_m is the number of channels used during spectral analysis. Note that in the current model, information sources are channel-specific, and thus separate channels can simultaneously occupy different states, i.e. $H_i^0(n)$ and $H_j^1(n)$ for $i \neq j$. The occurrence of individual inactive speech channels during temporal regions of active speech is due to such effects of speech production as harmonic valleys, and zeros of the vocal tract transfer function.

In determining speech presence probabilities, we are interested in calculating the posterior probability of active speech given a current observed speech spectral coefficient:

$$P\left(H_m^1 | x_m(n)\right). \quad (5)$$

Using a Bayesian approach, this value can be expressed as:

$$\begin{aligned} P\left(H_m^1 | x_m(n)\right) &= \frac{p\left(x_m(n)|H_m^1\right) P\left(H_m^1\right)}{\sum_{j=0}^{1} p\left(x_m(n)|H_m^j\right) P\left(H_m^j\right)} \\ &= \frac{\Lambda_m(n)}{1+\Lambda_m(n)}, \end{aligned} \quad (6)$$

where:

$$\Lambda_m(n) = \frac{P\left(H_m^1\right)}{P\left(H_m^0\right)} \frac{p\left(x_m(n)|H_m^1\right)}{p\left(x_m(n)|H_m^0\right)}. \quad (7)$$

The term $\Lambda_m(n)$ is often referred to as the generalized likelihood ratio (GLR). The conditional probabilities $p\left(x_m(n)|H_m^i\right)$ can be derived from statistical distributions of speech and noise spectral coefficients. In [1], the authors derive GLRs for the case of complex Gaussian noise spectral coefficients, and deterministic speech signals. In [2], the authors utilize complex Gaussian models for both speech and noise components, while in [20] results for the Laplacian case are presented.

The decoding method in this paper is robust with any statistical model. However, in this study we provide results for the method proposed in [2], arising from Gaussian models, due to its computational efficiency.

If independent complex Gaussian distributions are assumed for both noise and speech components, the GLR can be expressed as [2]:

$$\Lambda_m(n) = \frac{P\left(H_m^1\right)}{P\left(H_m^0\right)} \frac{1}{1+\xi_m(n)} \exp\left(\frac{\xi_m(n)}{1+\xi_m(n)}\gamma_m(n)\right). \quad (8)$$

Substitution of Eq. 8 into Eqs. 6-7 reveals the probability of active speech for each time-frequency location.

III. Improved Speech Presence Probabilities

Traditional SPPs (as described in Section II) are based solely on the current frame, and do not exploit the well-known temporal correlation present in spectral speech data. In this section, we present a novel algorithm for determining improved SPPs. The algorithm can operate in two modes: utilizing past and present observations, or utilizing past, present, and future observations, to fully take advantage of inter-frame correlation.

A. Interpreting SPPs as Observations of Channel-Specific 2-State Models

SPP masks derived from speech in favorable acoustic environments reveal reliable speech components which tend to occur in salient segments, which is due in part to the well-known temporal correlation of time-frequency speech data. Additionally, in favorable acoustic environments, SPPs tend towards binary masks, i.e. SPPs assume either $P\left(H_m^1\right)$ or 1. This can be shown by examining Equations 6 and 7 for extreme values of the *a priori* SNR. When the speech component is zero, and the additive noise is very small, $\xi_m\left(n\right)$=0. Furthermore:

$$\begin{aligned}\Lambda_m\left(n\right)|_{\xi_m(n)=0} &= \frac{P\left(H_m^1\right)}{P\left(H_m^0\right)} \\ \Rightarrow P\left(H_m^1|x_m\left(n\right)\right)|_{\xi_m(n)=0} &= P\left(H_m^1\right).\end{aligned} \tag{9}$$

Conversely, when the magnitude of the spectral speech component is much larger than that of the additive noise, then $\xi_m\left(n\right) \to \infty$. In this case:

$$\begin{aligned}\Lambda_m\left(n\right)|_{\xi_m(n)\to\infty} &= \infty \\ \Rightarrow P\left(H_m^1|x_m\left(n\right)\right)|_{\xi_m(n)\to\infty} &= 1.\end{aligned} \tag{10}$$

Thus, in the presence of very low background noise, SPP masks can be assumed binary. Furthermore, it is this binary mask that we interpret to contain "true" speech presence probabilities. However, the low noise case represents oracle information, and is not accessible in realistic speech processing systems; instead speech signals tend to be corrupted by higher levels of background noise. Therefore, in determining improved SPPs, we wish to estimate the underlying state, i.e. $P\left(H^1\right)$ or 1, of each spectro-temporal location given "noisy" standard SPP observations. Note that in order to simplify notation, we will refer to posterior probabilities obtained from traditional SPP methods as $\tau_m\left(n\right)$:

$$\tau_m\left(n\right) = P\left(H_m^1|x_m\left(n\right)\right). \tag{11}$$

By interpreting standard SPP masks as observations of channel-specific 2-state models, true binary masks can be estimated via traditional inference techniques. One such family of methods involves HMM-based decoding of noisy information [4].

B. HMM-Based Mask Decoding

Hidden Markov Models (HMMs) are characterized by steady-state, transitional, and observation statistics. In order to apply HMM-based inference techniques, these statistical parameters must be determined for each hidden two-state model, H_m^i. Transitional statistics are set to capture the temporal correlation present in spectral speech data. In this study, a_m^{ij} will denote the probability of transition between H_m^i and H_m^j, for $i,j \in \{0,1\},\ 1 \le m \le N_m$. The steady-state probability of state H_m^i can be obtained from transitional statistics as:

$$P\left(H_m^i\right) = \frac{\sum_{j=0,j\neq i}^{1} a_m^{ji}}{\sum_{g=0}^{1}\sum_{h=0,h\neq g}^{1} a_m^{gh}}. \tag{12}$$

Characterizing observation statistics for speech activity models H_m^i involves studying the effect of acoustic noise on corresponding speech presence probabilities. The distribution of observed reliability measures conditioned on underlying speech activity states is defined as:

$$b_m^i\left(\tau_m\left(n\right)\right) = p\left(\tau_m\left(n\right)|H_m^i\right),\ \text{for } 0 \le \tau_m\left(n\right) \le 1. \tag{13}$$

The relationship between additive acoustic noise in the spectral domain and the resulting inaccuracy of SPPs is difficult to express in closed form. Instead, we must rely on statistical tools to model the distribution of noisy SPPs about their underlying binary values, as a function of estimated additive acoustic noise.

It follows intuitively that the distribution of SPPs should reveal a global peak at the true binary value, and should display a monotonically decreasing probability density function (pdf) directly related to the distance from the underlying value. Considering these constraints, and due to their mathematical efficiency, we propose to model state-conditional observation distributions as raised cosine distributions:

$$\begin{aligned}&b_m^0\left(\tau_m\left(n\right)\right) = \\ &\quad\left(\frac{1}{1-P\left(H_m^1\right)}\right)\left(1+\phi_m^i\cos\left(\frac{\pi\left(\tau_m\left(n\right)-P\left(H_m^1\right)\right)}{1-P\left(H_m^1\right)}\right)\right), \\ &b_m^1\left(\tau_m\left(n\right)\right) = \\ &\quad\left(\frac{1}{1-P\left(H_m^1\right)}\right)\left(1+\phi_m^i\cos\left(\frac{\pi\left(1-\tau_m\left(n\right)\right)}{1-P\left(H_m^1\right)}\right)\right), \\ &\quad\text{for } p\left(H_m^1\right) \le \tau_m\left(n\right) \le 1.\end{aligned} \tag{14}$$

As can be observed from the expressions in Equation 14, the parameter $\phi_m^i \in [0,1]$ controls the effect of the sinusoidal component on the overall observation statistics. It follows intuitively that ϕ_m^i should be set to capture the estimated accuracy of observed channel-specific SPPs. That is, for clean spectro-temporal components, SPPs are determined with a high degree of accuracy, and thus observed close to their corresponding underlying binary states, i.e. $P\left(H_m^1\right)$ or 1. In this case, ϕ_m^i should be set close to 1. Conversely, for noisy spectro-temporal components, a higher degree of confusability is introduced into the estimation of SPPs, and observed SPPs

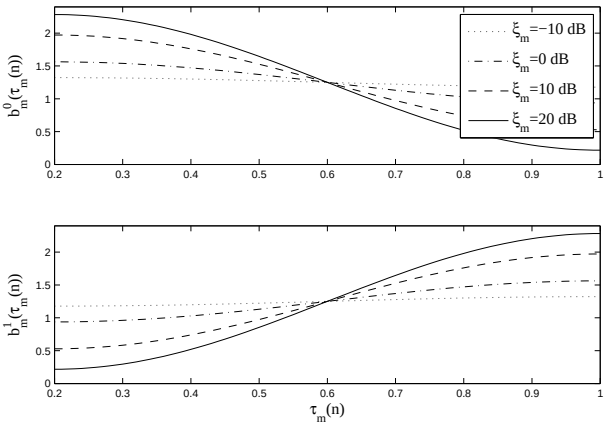

Fig. 1. Example probability distribution functions $b_m^0(\tau_m(n))$ and $b_m^1(\tau_m(n))$ for various values of $\xi_m(n)$. For this example, $P(H_m^1)$=0.2, and κ^i=1.0.

may vary from their underlying binary states. In this case, ϕ_m^i should be set close to 0.

Considering the relationship between additive noise and the statistical parameter ϕ_m^i, we propose the parameter to be a function of the *a priori* SNR, ξ_m, of the given spectro-temporal component. Specifically, we propose ϕ_m^i to be determined in time as:

$$\phi_m^i(n) = \left(\frac{\xi_m(n)}{\xi_m(n)+\kappa^i}\right)^2. \tag{15}$$

Here, κ^i is an empirically tuned parameter. Figure 1 provides example probability distribution functions $b_m^0(\tau_m(n))$ and $b_m^1(\tau_m(n))$ for various values of $\xi_m(n)$. For this example, $p(H_m^1)$=0.2. As can be interpreted, when the *a priori* SNR is high (e.g. ξ_m=20 dB), the corresponding distribution decays rapidly. However, when the *a priori* SNR is low (e.g. ξ=$-$10 dB), the corresponding distribution decreases slowly, similar to a uniform pdf.

In this paper, we propose improved speech presence probabilities, $\hat{\tau}_m(n)$, based on the set of traditional SPPs, $\{\tau_m(1),\ldots,\tau_m(n)\}$, by exploiting the temporal correlation present in spectral speech data. Once underlying channel-specific models are parameterized, HMM-based decoding can be utilized to determine the minimum mean-square error (MMSE) estimate of the true binary SPP mask. The decoded SPP corresponding to the observed probability $\tau_m(n)$ is determined via the forward algorithm as [4]:

$$\begin{aligned}\hat{\tau}_m(n) &= P\left(H_m^1(n)\,|\tau_m(1),\ldots,\tau_m(n)\right) \\ &= \frac{\alpha_m^1(n)}{\alpha_m^0(n)+\alpha_m^1(n)}\end{aligned} \tag{16}$$

where $\alpha_m^i(n)$ is the forward variable for channel m, corresponding to state i, at time index n. Forward variables convey the probability of the current observation occupying state i, given past observations:

$$\alpha_m^i(n) = P\left(H_m^i(n)\,|x_m(1),\ldots,x_m(n)\right). \tag{17}$$

The forward variables $\alpha_m^i(n)$ and can be determined recursively as:

$$\alpha_m^i(n) = \begin{cases} \left[\sum_{j=0}^1 a^{ji}\alpha_m^j(n-1)\right] b_m^i(\tau_m(n)), \\ \qquad\qquad \text{if } n>1 \\ 1, \text{ else} \end{cases} \tag{18}$$

Thus, Equation 16 provides improved speech presence probabilities, given current and past standard SPPs.

C. Incorporating Future Observations

For applications in which slight delays are acceptable, improved SPPs can be determined by incorporating future observations via the forward-backward algorithm [4]. In this case, similar to Eq. 16, the decoded speech presence probability is:

$$\begin{aligned}\hat{\tau}_m(n) &= P\left(H_m^1(n)\,|\tau_m(1),\ldots,\tau_m(n+N_{LA})\right) \\ &= \frac{\alpha_m^1(n)\,\beta_m^1(n,0)}{\alpha_m^1(n)\,\beta_m^1(n,0)+\alpha_m^0(n)\,\beta_m^0(n,0)}\end{aligned} \tag{19}$$

where N_{LA} is the total number of look-ahead frames utilized. Here, $\beta_m^i(n,k)$, the backward variable, differs from traditional notation in that it is a function of two parameters. This is due to the generally time-sensitive nature of tasks such as speech enhancement, which forces the recursive calculation of backward variables to be re-initialized for each time index n. The backward variable $\beta_m^i(n,k)$ conveys the probability of channel m occupying state i at time index $n+k$, during recursive calculations ultimately required for time index n:

$$\begin{aligned}&\beta_m^i(n,k) \\ &= P\left(H_m^i(n+k)\,|x_m(n+k),\ldots,x_m(n+N_{LA})\right).\end{aligned} \tag{20}$$

The backward variable $\beta_m^i(n,k)$ and can be determined recursively as:

$$\beta_m^i(n,k) = \begin{cases} \sum_{j=0}^1 a^{ij}\beta_m^j(n,k+1)\, b_m^j(\tau_m(n+k+1)), \\ \qquad\qquad \text{if } k<N_{LA} \\ 1, \text{ else} \end{cases}. \tag{21}$$

In this way, future observations can be exploited to improve the estimation of standard SPPs.

D. Complexity Analysis

A well known downside to HMM-based processing is the induced computational load. This is especially problematic for speech applications, which can be delay-sensitive and/or resource constrained. However, due to the small size of the underlying model used during estimation of improved SPPs, the induced complexity is relatively small.

Table I provides operations required by the proposed algorithms for determining SPPs. The traditional method from [2], and the improved method from [12] are included for reference. It can be observed from Table I that the additional number of operations required, as compared to [2], is relatively low, making it an attractive option for resource-constrained applications. Furthermore, methods proposed in this paper induce a significantly smaller computational load than that of [12]. In Table I, number of operations is given per frame and per frequency channel. Note that the induced load of the fast Fourier transform (FFT) is not included, but is known to be of order $O\left(N_m \log\left(N_m\right)\right)$.

IV. Experimental Results

A. The Database for Speech with Additive Noise

As previously discussed, robust algorithms in this study are designed for the additive noise case. During the experimental procedure, 20 randomly selected sentences from the TIMIT database were used. Along with stationary white noise, 2 non-stationary noise types were studied, namely restaurant and babble. Speech and noise signals were mixed according to the FANT algorithm [9] at SNRs between 20 and 0 dB.

B. Accuracy of Improved SPPs

Figure 2 presents illustrative examples of SPP masks determined by various methods. Panel (**a**) provides the clean speech signal "She had your dark suit in greasy wash water all year" by a female speaker. Panel (**b**) shows the SPP mask determined according to [12] from a corresponding signal corrupted by airport noise at 15 dB SNR. Panels (**c**) and (**d**) provide proposed SPP masks according to Eq. 16 and Eq. 19 (N_{LA}=2), respectively. As can be observed in panel (**b**), the algorithm proposed in [12] results in a high false alarm rate. The proposed mask in panel (**c**) significantly reduces the false alarm rate, and manages to detect individual harmonics. Incorporating future observations in (**d**) results in a smoothed mask, wherein detected speech regions are presented in more salient segments.

To grade the accuracy of SPPs[1], we utilize pointwise Kullback-Leibler (KL) distances [5] between masks obtained from noisy speech and those obtained from corresponding "oracle" clean speech. The KL distance is suitable since it is commonly used to compare statistical distributions, and time- and frequency-specific SPPs are observations of individual pdfs. The mean pointwise KL distance between masks is given by:

$$\bar{D}\left(\tau^{orc}||\hat{\tau}\right) = \frac{1}{N_t N_m} \sum_{n=1}^{N_t} \sum_{m=1}^{N_m} \tau_m^{orc}\left(n\right) \log\left(\frac{\tau_m^{orc}\left(n\right)}{\hat{\tau}_m\left(n\right)}\right), \tag{22}$$

where τ^{orc} refers to the oracle mask, and N_t denotes the length of the given sound file in frames. Note that oracle masks are determined by adding artificial white noise to clean speech at ≈ 40 dB, and following steps outlined in Section II.

[1]Application of algorithms described in III requires certain numerical parameters, which are included in Table II

TABLE III
Mean Pointwise Kullback-Leibler (KL) Distance for SPP Masks From Oracle Masks, In Bits

SPP Method	SNR (dB) 20	15	10	5	0
Non-stationary Restaurant Noise					
from [12]	0.13	0.20	0.34	0.57	0.86
Eq. 16	0.10	0.17	0.28	0.42	0.54
Eq. 19 (N_{LA}=2)	0.09	0.17	0.29	0.44	0.60
Non-stationary Babble Noise					
from [12]	0.13	0.23	0.41	0.67	0.95
Eq. 16	0.12	0.21	0.32	0.45	0.58
Eq. 19 (N_{LA}=2)	0.11	0.21	0.34	0.50	0.65
Stationary White Noise					
from [12]	0.52	1.02	1.63	2.27	2.83
Eq. 16	0.33	0.50	0.66	0.79	0.89
Eq. 19 (N_{LA}=2)	0.34	0.56	0.78	0.96	1.10
Average					
from [12]	0.26	0.48	0.79	1.17	1.54
Eq. 16	0.18	0.29	0.42	0.55	0.67
Eq. 19 (N_{LA}=2)	0.18	0.31	0.47	0.64	0.79

Table III provides mean pointwise KL distances for SPP masks obtained for non-stationary colored noise, from corresponding oracle masks. As reference, results for SPPs from [12] are included. Note that similar to the proposed method in this paper, the work in [12] exploits past observations. However, while [12] combines current and past observations in a somewhat heuristic manner, the proposed method utilizes a statistical framework to find the MMSE estimate given the HMM framework. As can be concluded from Table III, the proposed speech presence probabilities provide a significant increase in mask accuracy in terms of the pointwise KL distance. It should be noted that for certain cases, the KL distance increases with the inclusion of look-ahead frames. Using look-ahead frames tends to result in high SPPs occurring in salient segments due to a higher degree of HMM-based smoothing. It, in turn, provides lower missed detection rates, while increasing false alarms.

C. Soft-Decision Speech Enhancement

We apply the proposed method for determining improved SPPs to minimum mean-square error (MMSE) soft-decision noise speech enhancement to further grade its performance. Traditional statistical noise suppression rules for speech operate under the assumption of speech presence throughout spectro-temporal locations. However, this assumption is untrue both for time periods of inactive speech, and for frequency channels corresponding to harmonic valleys or vocal tract zeros during active speech. Instead, utilizing speech presence probability during noise suppression takes into account speech presence uncertainty, and leads to soft-decision speech enhancement.

Speech presence probabilities can be integrated into noise suppression techniques by deriving a MMSE estimate of speech spectral coefficients [19]:

TABLE I

REQUIRED OPERATIONS FOR PROPOSED SPPS: NUMBERS OF OPERATIONS ARE GIVEN PER FRAME AND PER FREQUENCY CHANNEL. NOTE THAT THE INDUCED LOAD OF THE FAST FOURIER TRANSFORM (FFT) IS NOT INCLUDED, BUT IS KNOWN TO BE OF ORDER $O\left(N_m \log\left(N_m\right)\right)$.

SPP Method	$+$	$\times$	$\div$	cos	exp	log
Traditional SPPs[2]	4	6	3	0	1	0
Improved SPPs from [12]	40	45	8	0	0	4
Proposed SPPs (Eq. 16)	12	20	5	2	1	0
Proposed SPPs (Eq. 19)	$12+2N_{LA}$	$20+8N_{LA}$	5	2	1	0

TABLE II

NUMERICAL PARAMETERS FOR PROPOSED SPPS DESCRIBED IN SEC. III, FOR SPEECH ENHANCEMENT AND AUTOMATIC SPEECH RECOGNITION (ASR)

Parameter	Enhancement	ASR	Description
N_m	257	257	number of channels during spectral analysis
a_m^{01}	0.05	0.20	HMM transitional probability, $H_m^0 \rightarrow H_m^1$
a_m^{10}	0.10	0.20	HMM transitional probability, $H_m^1 \rightarrow H_m^0$
κ^0	1.0	1.0	used by observation probability, $b_m^0\left(\tau_m\left(n\right)\right)$
κ^1	10.0	7.0	used by observation probability, $b_m^1\left(\tau_m\left(n\right)\right)$

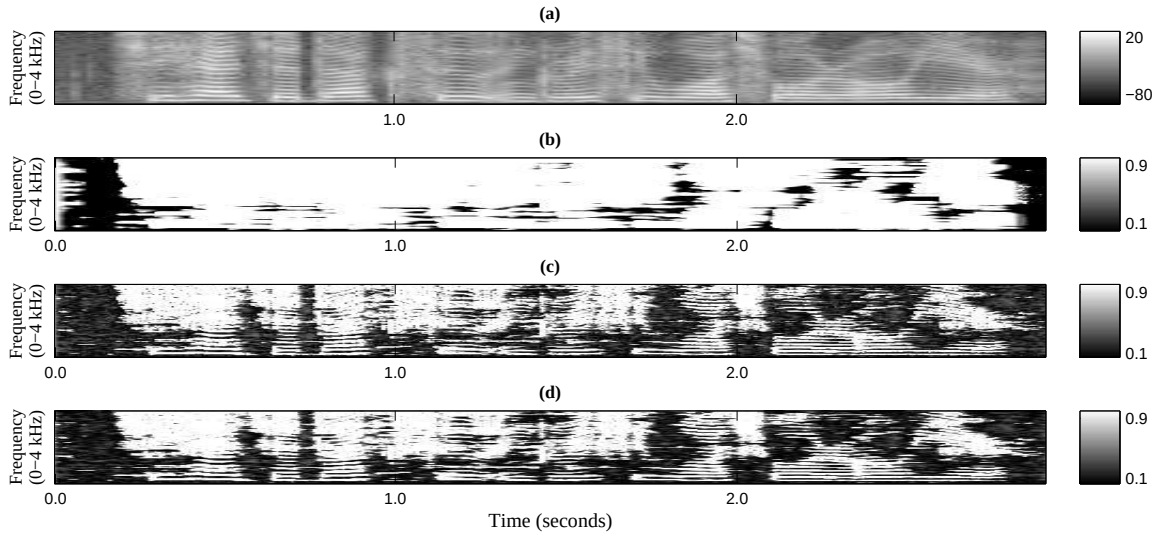

Fig. 2. Illustrative Examples of SPP Masks Determined by Various Methods: Panel (**a**) provides the clean speech signal "She had your dark suit in greasy wash water all year" spoken by a female. Panel (**b**) shows the SPP mask determined according to [12] from a corresponding signal corrupted by airport noise at 15 dB SNR. Panels (**c**) and (**d**) provide proposed SPP masks according to Eq. 16 and Eq. 19 (N_{LA}=2), respectively.

$$\begin{aligned}\hat{s}_m\left(n\right) &= E[s_m\left(n\right)|x_m\left(n\right)] \\ &= E[s_m\left(n\right)|x_m\left(n\right), H_m^0]P\left(H_m^0|x_m\left(n\right)\right) \\ &+ E[s_m\left(n\right)|x_m\left(n\right), H_m^1]P\left(H_m^1|x_m\left(n\right)\right).\end{aligned} \tag{23}$$

It follows intuitively that the expected value of speech coefficients, conditioned on an inactive speech state, is zero:

$$E[s_m\left(n\right)|x_m\left(n\right), H_m^0] = 0. \tag{24}$$

The MMSE noise suppression rule is derived by integrating a gain factor in Equation 23 [1], [2], [3]:

$$\hat{s}_m\left(n\right) = P\left(H_m^1|x_m\left(n\right)\right) G\left(x_m\left(n\right)\right) x_m\left(n\right), \tag{25}$$

where $G\left(x_m\left(n\right)\right)$ is generalized as:

$$G\left(x_m\left(n\right)\right) = \frac{E\left[s_m\left(n\right)|x_m\left(n\right), H_m^1\right]}{x_m\left(n\right)}. \tag{26}$$

In this study, we utilize the optimally modified log-spectral amplitude (OM-LSA) estimator proposed in [14] as an illustrative example. Additionally, noise estimation was performed according to [12]. Code for the previously discussed algorithms was obtained from [23]. It should be noted that the numerous parameters defined by [12] were kept as in [23].

The performance of proposed SPP masks during speech enhancement was tested by analyzing their ability to minimize acoustic noise leakage (NL) while maintaining low speech distortion. Similar to the experimental procedure in [22], we define noise leakage as the percentage of energy corresponding

to time-frequency bins deemed as inactive speech by the oracle mask, which passes unsuppressed by SPP masks:

$$NL = 100\frac{\sum_{n=1}^{N_t}\sum_{m=1}^{N_m}\max\left[\hat{\tau}_m\left(n\right)-\tau_m^{orc}\left(n\right),0\right]\left|x_m\left(n\right)\right|^2}{\sum_{n=1}^{N_t}\sum_{m=1}^{N_m}\max\left[1-\tau_m^{orc}\left(n\right),0\right]\left|x_m\left(n\right)\right|^2}. \tag{27}$$

Conversely, we define speech distortion (SD) as the percentage of energy corresponding to active speech bins, which is distorted by SPP masks:

$$SD = 100\frac{\sum_{n=1}^{N_t}\sum_{m=1}^{N_m}\max\left[\tau_m^{orc}\left(n\right)-\hat{\tau}_m\left(n\right),0\right]\left|x_m\left(n\right)\right|^2}{\sum_{n=1}^{N_t}\sum_{m=1}^{N_m}\max\left[\tau_m^{orc}\left(n\right),0\right]\left|x_m\left(n\right)\right|^2}. \tag{28}$$

There exists a natural trade-off between NL and SD, i.e. as a greater percentage of acoustic noise is suppressed, more distortion to the underlying speech signal is generally expected.

Table IV provides results for speech distortion and noise leakage for proposed SPP masks during soft-decision speech enhancement. The state-of-the-art method from [12] is included as reference. Proposed SPPs (Eq. 16) are shown to provide low SD, while significantly decreasing the NL for most conditions. Integrating future observations in proposed SPPs (Eq. 19, N_{LA}=1, 2) generally results in a decrease in both SD and NL. For non-stationary noise, particularly at low SNRs, the speech distortion observed for the proposed mask estimation technique is substantially greater than that observed for [12]. This is most probably due to the dependency of the proposed algorithm on the *a priori* SNR, which is difficult to estimate in non-stationary conditions.

The performance of proposed SPP masks was also assessed by traditional speech distortion metrics. Table V provides improvements in segmental SNR (ΔSSNR) for enhanced signals using proposed SPPs. It can be concluded that proposed method generally provides increased improvement over those presented in [12], although these changes may not be generally perceptually significant.

TABLE V
IMPROVEMENTS IN SEGMENTAL SNR ($\Delta SSNR$) FOR SOFT-DECISION SPEECH ENHANCEMENT USING SPP MASKS

	SNR (dB)				
SPP Method	20	15	10	5	0
Non-stationary Restaurant Noise					
from [12]	3.12	3.99	4.73	5.33	5.83
Eq. 16	3.20	4.21	4.97	5.52	6.02
Eq. 19 (N_{LA}=1)	3.30	4.29	5.03	5.53	5.92
Non-stationary Babble Noise					
from [12]	4.34	5.34	6.33	7.35	8.43
Eq. 16	4.34	5.47	6.51	7.59	8.60
Eq. 19 (N_{LA}=1)	4.51	5.61	6.59	7.60	8.57
Stationary White Noise					
from [12]	3.96	5.12	6.39	7.66	8.79
Eq. 16	3.76	5.05	6.41	7.83	9.15
Eq. 19 (N_{LA}=1)	4.07	5.32	6.66	8.03	9.31
Average					
from [12]	3.81	4.82	5.82	6.78	7.68
Eq. 16	3.77	4.91	5.96	6.97	7.92
Eq. 19 (N_{LA}=1)	3.96	5.08	6.09	7.04	7.91

TABLE VI
WORD-ACCURACY RESULTS FOR ASR USING FRONT-END SOFT-DECISION SPEECH ENHANCEMENT WITH SPP MASKS

Front-End	SNR (dB)				
Noise Suppression	20	15	10	5	0
Test Set A					
None	97.32	91.70	69.00	37.50	15.39
OM-LSA [12]	95.44	91.58	82.86	66.65	42.38
Eq. 29 (N_{LA}=2)	97.40	94.80	87.93	72.22	43.73
Test Set B					
None	97.44	88.91	61.73	31.72	14.59
OM-LSA [12]	94.47	89.34	80.48	63.93	39.77
Eq. 29 (N_{LA}=2)	96.80	92.57	83.91	65.38	35.54
Average					
None	97.38	90.31	65.36	34.61	14.99
OM-LSA [12]	94.96	90.46	81.67	65.29	41.68
Eq. 29 (N_{LA}=2)	97.10	93.69	85.92	68.80	39.64

D. Noise Robust Automatic Speech Recognition

Soft-decision speech enhancement was applied as a front-end noise robust technique for automatic speech recognition to assess the quality of resulting signals. Front end feature extraction included 13 MFCCs and log-energy, along with first and second derivatives. The HMM-based recognizer utilized 16-state, 3-mixture word models. Test Sets A and B of the Aurora-2 database, comprised of connected digit utterances, were used during experimentation.

Enhanced speech spectral coefficients were determined by MMSE estimation:

$$\begin{aligned}\hat{s}_m\left(n\right) &= E[s_m\left(n\right)|x_m\left(n\right)] \\ &= E[s_m\left(n\right)|x_m\left(n\right),H_m^1]P\left(H_m^1|x_m\left(n\right)\right) \\ &\approx x_m\left(n\right)P\left(H_m^1|x_m\left(n\right)\right)\end{aligned} \tag{29}$$

That is, the gain function $G\left(x_m\left(n\right)\right)$, from Eq. 25, was excluded to avoid over-suppression of discriminative speech information important for recognition. Table VI provides word-accuracy results for ASR using front-end soft-decision speech enhancement with SPP masks. As reference, results for the OM-LSA speech enhancement system of [12] are included, as well as results for the baseline system which includes no front-end noise suppression. As can be observed in Table VI, soft-decision enhancement using proposed SPP masks provides improved ASR performance relative to the system in [12].

V. CONCLUSIONS

In this paper we have presented a framework for determining improved SPPs using HMM-based inference. We model spectro-temporal data as observations from channel-specific two-state models, and apply HMM-based decoding to estimate true posterior probabilities.

We illustrate the effectiveness of the proposed framework by applying it to soft-decision speech enhancement. The use of proposed SPPs is shown to provide significant improvements in mask accuracy over the state-of-the-art SPP method in [12]

TABLE IV

SPEECH DISTORTION (SD) AND NOISE LEAKAGE (NL) RESULTS FOR PROPOSED SPP MASKS DURING SOFT-DECISION SPEECH ENHANCEMENT: PROPOSED TECHNIQUES SHOW LOW SD WHILE PROVIDING SIGNIFICANTLY REDUCED NL, RELATIVE TO [12], FOR MOST NOISE CONDITIONS.

	20 dB		15 dB		10 dB		5 dB		0 dB	
SPP Method	SD (%)	NL (%)	SD (%)	NL (%)	SD (%)	NL (%)	SD (%)	NL (%)	SD (%)	NL (%)
Non-stationary Restaurant Noise										
from [12]	0.04	84.09	0.10	74.89	0.37	62.97	1.08	51.41	3.03	42.57
Eq. 16	0.23	49.45	0.64	34.15	1.64	23.17	4.05	18.03	8.93	15.43
Eq. 19 (N_{LA}=1)	0.16	50.16	0.48	33.71	1.30	20.58	3.45	14.76	8.01	12.37
Eq. 19 (N_{LA}=2)	0.14	54.60	0.40	38.95	1.10	25.27	2.98	18.25	7.05	14.99
Non-stationary Babble Noise										
from [12]	0.05	83.62	0.23	72.98	0.87	60.38	2.25	49.05	5.42	41.05
Eq. 16	0.39	47.41	1.14	32.95	2.90	22.69	6.30	16.34	13.07	13.54
Eq. 19 (N_{LA}=1)	0.29	47.57	0.92	32.01	2.52	19.75	5.61	13.07	12.07	10.23
Eq. 19 (N_{LA}=2)	0.25	52.16	0.80	36.93	2.23	24.31	4.97	16.62	10.80	12.89
Stationary White Noise										
from [12]	0.14	44.27	0.52	26.81	1.54	14.80	4.33	6.76	11.19	2.75
Eq. 16	0.50	10.97	1.34	7.49	3.27	5.57	7.35	4.41	15.05	3.83
Eq. 19 (N_{LA}=1)	0.38	6.36	1.08	3.14	2.78	1.76	6.57	1.09	14.09	0.83
Eq. 19 (N_{LA}=2)	0.33	9.48	0.95	4.72	2.48	2.64	5.94	1.69	12.91	1.31
Average										
from [12]	0.08	70.66	0.28	58.23	0.92	46.05	2.55	35.74	6.55	28.79
Eq. 16	0.38	35.94	1.04	24.86	2.60	17.00	5.90	12.93	12.35	10.93
Eq. 19 (N_{LA}=1)	0.28	34.70	0.83	22.95	2.20	14.03	5.21	9.64	11.39	7.81
Eq. 19 (N_{LA}=2)	0.24	38.74	0.72	26.87	1.94	17.41	4.63	12.18	10.25	9.73

in terms of the mean pointwise KL distance. When applied to the task of soft-decision speech enhancement, proposed method is shown to improve performance in terms of segmental SNR. Closer analysis of enhanced speech signals reveals a significant decrease in noise leakage, while speech distortion is observed to increase. For high SNRs, particularly in stationary noise conditions, the increase in SD can be concluded to be minimal. At lower SNRs, however, the SD is seen to increase by fairly substantial amounts. When applied as a front-end noise robust method for ASR, soft-decision enhancement using proposed SPP masks provides improved recognition performance relative to [12], for favorable acoustic conditions. Furthermore, the 2-state configuration of underlying HMMs results in a relatively small increase in complexity, making the proposed method attractive for resource-constrained scenarios.

SPP masks serve as pivotal tools in noise robust speech processing systems. Improved SPPs proposed in this paper can be utilized in applications such as voice activity detection (VAD) and pitch estimation/tracking. Such topics will be the focus of future work.

REFERENCES

[1] R. J. McAuley and M. L. Malpass, *Speech Enhancement Using a Soft-Decision Noise Suppression Filter*, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 28, No. 2, pp. 137-145, 1980.

[2] Y. Ephraim and D. Malah, *Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator*, Vol. 32, pp. 1109-1121, 1984

[3] Y. Ephraim and D. Malah, *Speech Enhancement Using a Minimum Mean-Square Log-Spectral Amplitude Estimator*, IEEE Trans. on Acoustics, Speech ,and Signal Processing, Vol. 33, No. 2, pp. 443-445, 1985.

[4] 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.

[5] T. Cover and J. Thomas,*Elements of Information Theory*, Wiley, 1991.

[6] J. Sohn, N. S. Kim, and W. Sung, *A Statistical Model-Based Voice Activity Detection*, IEEE Signal Processing Letters, Vol. 6, No. 1, pp. 1-3, 1999.

[7] D. Malah, R. V. Cox, and A. J. Accardi, *Tracking Speech Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments*, ICASSP, Vol. 2, pp. 789-792, 1999.

[8] D. Pearce, *Enabling New Speech Driven Services For Mobile Devices: An Overview of the ETSI Standards Activities for Distributed Speech Recognition Front-Ends*, AVIOS 2000: Speech Appl. Conf., Vol. 5, pp. 1-6, May 2000.

[9] H.-G. Hirsch and D. Pearce, *The AURORA Experimental Framework For The Performance Evaluation of Speech Recognition Systems Under Noisy Conditions*, ASR2000 - Automatic Speech Recognition: Challenges for the new Millenium, 2000.

[10] P. Renevey and A. Drygajlo, *Entropy Based Voice Activity Detection in Very Noisy Conditions*, Eurospeech, pp. 1887-1890, 2001.

[11] R. Martin, *Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics*, IEEE Trans. Speech and Audio Processing, Vol. 9, No. 5, pp. 504-512, 2001.

[12] I. Cohen and B. Berdugo, *Speech Enhancement for Non-Stationary Noise Environments*,Signal Processing, Vol. 81, Issue 11, pp. 2403-2418, 2001.

[13] ITU-T Rec. P.862 Perceptual Evaluation of Speech Quality (PESQ)

[14] I. Cohen, *Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator*, IEEE Signal Processing Letters, Vol. 9, No. 4, pp. 113-116, 2002.

[15] I. Cohen, *Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging*, IEEE Trans. Speech and Audio Processing, Vol. 11, No. 5, pp. 466-475, 2003.

[16] R. Martin and C. Breithaupt, *Speech Enhancement in the DFT Domain Using Laplacian Speech Priors*, Workshop on Acoustic Echo and Noise Control (IWAENC), pp. 87-90, 2003.

[17] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, *Efficient Voice Activity Detection Algorithms Using Long-Term Speech Information*, Speech Communication, Vol. 42, Issues 3-3, pp. 271-287, 2004.

[18] G. Almpanidis and C. Kotropoulos, *Voice Activity Detection with Generalized Gamma Distribution*, ICME, pp. 961-964, 2006.

[19] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, New York: Wiley, 2006.

[20] B. Chen and P. Loizou, *A Laplacian-based MMSE Estimator for Speech Enhancement*, Speech Communication, Vol. 49, Issue 2, pp. 134-143, 2007

[21] ETSI Standard Doc., *Speech Processing, Transmission, and Quality*

Aspects (STQ); Distributed Speech Recognition; Front-end Feature Extraction Algorithms; Compression Algorithms, ETSI ES 202 050 v1.1.1 (2007-10).

[22] T. Gerkmann, C. Breithaupt, and R. Martin, *Improved A Posteriori Speech Presence Probability Estimation Based on a Likelihood Ratio with Fixed Priors*, IEEE Trans. on Audio, Speech, and Language Processing, Vol. 16, No. 5, pp. 910-919, 2008.

[23] http://webee.technion.ac.il/Sites/People/IsraelCohen/