We present an algorithm for identifying the location of sibilant phones in noisy speech. Our algorithm does not attempt to identify sibilant onsets and offsets directly but instead detects a sustained increase in power over the entire duration of a sibilant phone. The normalized estimate of the sibilant power in each of 14 frequency bands forms the input to two Gaussian mixture models that are trained on sibilant and non-sibilant frames respectively. The likelihood ratio of the two models is then used to classify each frame. We evaluate the performance of our algorithm on the TIMIT database and demonstrate that the classification accuracy is over 80% at 0 dB signal to noise ratio for additive white noise.

Index Terms: sibilant speech, spectrographic mask estimation, speech classification, speech segregation

1. INTRODUCTION

Sibilant speech accounts for a large fraction of aperiodic high frequency speech energy and, in English, comprises the fricatives /s/, /ʃ/, /z/ and /ʒ/ and the affricatives /tʃ/ and /dʒ/. The importance of aperiodic speech detection has been highlighted in several speech processing applications. In [1], it was shown that enhancing noisy unvoiced speech plays a greater role in achieving accurate speech recognition than enhancing voiced speech. An increasing interest in unvoiced speech detection has also emerged for speech enhancement, especially for spectrographic mask estimation [2, 3], where most previous approaches have focused on voiced speech segregation [4, 5].

Robust approaches to identify and segregate unvoiced speech have been proposed in [2] and [3] to improve the performance of binary-mask-based speech enhancement algorithms. Both approaches implement as a first step, an auditory segmentation employing a multi-scale analysis of event onsets and offsets. Although the segment classification is conducted in a different way in each approach, both rely on the correct identification of the fundamental frequency.

In this paper, we present a sibilant detection algorithm robust to high levels of noise for wide-band speech that operates in the frequency domain and that does not rely on voicing detection. Rather than identifying explicit sibilant onsets and offsets, a sustained increase in energy during the sibilant is instead detected. Under the hypothesis of a sibilant presence within a time-frame, its mean power in each frequency band is estimated using a maximum likelihood approach. This information is sent to a classifier which discriminates sibilant from non-sibilant time frames.

2. PROPOSED METHOD

Following [6], we assume that the short-time Fourier transform (STFT) coefficients of speech and noise can be modeled as statistically independent complex Gaussian random variables. Given a noisy speech signal, the power, $X_{k,i}$, in a time-frequency STFT bin is therefore distributed as

$$p(x_{k,i}) = \frac{1}{\mu_{k,i}} \exp \left( -\frac{x_{k,i}}{\mu_{k,i}} \right)$$  \hspace{1cm} (1)

where $k$ and $i$ are the frequency and time-frame indices and $\mu_{k,i}$ is the mean power.

Figure 1: Power spectral density (PSD) at 5 kHz versus time of a speech segment containing the sibilant phone /ʃ/ using a Hamming analysis window of 3.6 ms duration with 75% overlap. The time origin represents the center of the sibilant phone.

Fig. 1 shows the time-variation of power at 5 kHz for a noisy speech example containing the phone /ʃ/. We can divide the time interval into three segments as shown: a central segment $S$ that encompasses the sibilant and two surrounding intervals, $N1$ and $N2$, that contain no sibilant energy. We assume the mean power of the
speech to be constant over $S$ and that of the noise to be constant over the entire interval $N_1 + S + N_2$, giving
\[ \mu_{k,N_1} = \ldots \]
one trained on non-sibilant speech
and the other on sibilant speech. The probability that a

\begin{equation}
L_k = \sum_{i \in S} \left( -\ln(a_k + b_k) - \frac{x_{k,i}}{a_k + b_k} \right) \\
+ \sum_{i \in N_1,N_2} \left( -\ln(a_k) - \frac{x_{k,i}}{a_k} \right) \tag{2}
\end{equation}

2.1. Sibilant Speech Energy Estimation

By maximizing the log-likelihood in (2), the sibilant mean, $b_k$, and the noise mean, $a_k$, can be estimated if the exact time and duration of the sibilant phone are known. However, the duration of an actual sibilant is unknown and varies in each case. Fig. 2 shows the sibilant duration distribution in the TIMIT training set [7]. We observe that 74% of sibilants durations lie within 60 and 130 ms. Therefore, if $t = 0$ represents the center of a sibilant $|t| < 30$ ms has a high probability of lying within the sibilant while the region $|t| > 65$ ms has a high probability of lying outside the sibilant. To account for this, we apply a weighting function, $w_i$, to the time frames when calculating the log-likelihood that reduces the contribution of the transition region 30 ms $< |t| < 65$ ms as shown in Fig. 4. The weighted log-likelihood can now be expressed as

\[ \tilde{L}_k = \sum_{i \in S} w_i \left( -\ln(a_k + b_k) - \frac{x_{k,i}}{a_k + b_k} \right) \\
+ \sum_{i \in N_1,N_2} w_i \left( -\ln(a_k) - \frac{x_{k,i}}{a_k} \right) \tag{3}
\]

We maximise the value of the log-probability with respect to $a_k$ and $b_k$ by setting the partial derivatives to zero

\[ 0 = \frac{\partial \tilde{L}_k}{\partial a_k} = \sum_{i \in S} w_i \left( -\frac{1}{a_k + b_k} - \frac{x_{k,i}}{(a_k + b_k)^2} \right) \\
+ \sum_{i \in N_1,N_2} w_i \left( -\frac{1}{a_k} - \frac{x_{k,i}}{a_k^2} \right) \tag{4}
\]

\[ 0 = \frac{\partial \tilde{L}_k}{\partial b_k} = \sum_{i \in S} w_i \left( -\frac{1}{a_k + b_k} - \frac{x_{k,i}}{(a_k + b_k)^2} \right) \\
+ \sum_{i \in N_1,N_2} w_i \left( -\frac{1}{a_k} - \frac{x_{k,i}}{a_k^2} \right) \tag{5}
\]

from which we can estimate the mean noise energy, $a_k$, and the mean sibilant energy, $b_k$, as

\[ \hat{a}_k = \frac{\sum_{i \in N_1,N_2} w_i x_{k,i}}{\sum_{i \in N_1,N_2} w_i} \tag{6} \]

\[ \hat{b}_k = \frac{\sum_{i \in S} w_i x_{k,i}}{\sum_{i \in S} w_i} - \hat{a}_k \tag{7} \]

Under the hypothesis that time-frame $n$ lies at the center of a fixed-length sibilant phone, we can estimate the mean sibilant power in frequency bin $k$ using (7). We denote this estimate as $\hat{b}_{k,n}$, where the index $n$ represents the time-frame considered to be the center of segment $S$. Fig. 3(a) shows the PSD waveform of $\hat{b}_{k,n}$ for the \textipa{/j}/ sibilant example shown in Fig. 1. We see that it reaches a maximum when $n$ lies near the centre of the phone and becomes negative either side of the phone when region $N_1$ or $N_2$ overlaps significantly with the true sibilant.

2.2. Maximum filter and normalization

The quantity $\hat{b}_{k,n}$ from (7) will give a reliable estimate of sibilant power near the center of a sibilant phone and also in signal regions where no sibilant is present. However, as can be seen in Fig. 3(a), the estimate of sibilant power is less accurate in frames near the sibilant boundary. To counter this effect, we apply a maximum filter to the sibilant power estimate

\[ \tilde{b}_{k,n} = \frac{\max_{|m-n|<W/2} \hat{b}_{k,m}}{W} \tag{8} \]

where $W$, the filter support, represents the minimum sibilant duration. Fig. 3(b) shows the filter output, $\tilde{b}_{k,n}$, using $W = 30$ ms and we observe that the estimated $\tilde{b}_{k,n}$ remains at a high level for most of the sibilant duration.

To make the estimate independent of the overall speech level, the estimated sibilant mean power within each frame is normalized to give

\[ \hat{b}_{k,n} = \frac{\tilde{b}_{k,n}}{\frac{1}{K} \sum_{k=1}^{K} |\hat{b}_{k,n}|} \tag{9} \]

The absolute value is used because as seen in Fig. 3(b), $\hat{b}_{k,n}$ can be negative when the sibilant occupies a region that was assumed to be noise.

2.3. Gaussian Mixture Model

For each frame, the normalized sibilant power spectrum, $\hat{b}_{k,n}$ for $k \in [1, K]$, forms the input to two Gaussian mixture models (GMMs): one trained on non-sibilant speech and the other on sibilant speech. The probability that a
time frame contains a sibilant phone is calculated from the likelihood ratio of the two GMMs.

3. EXPERIMENTS

The sibilant detector described in this paper includes a number of algorithm parameters whose values were determined using the training set of the TIMIT database [7], which includes phonetic transcription. The STFT used a Hamming analysis window of 3.6 ms duration with 75% overlap. The relatively short analysis window provides a high time resolution and a frequency resolution that is able to characterize the sibilant power spectrum without resolving pitch harmonics. The power spectrum of each frame was interpolated using triangular filters to give 14 frequency bins whose centers are uniformly spaced from 1.5 kHz to 8 kHz.

The sibilant duration, $S$, as well as the duration of $N1$ and $N2$ need to be fixed in order to estimate the mean sibilant energy, $b_{k,n}$, from equation (7). We evaluated a range of fixed widths for $S$ as well as a variable width approach in which (3) was maximized with respect to the phone boundaries in addition to the powers $a_k$ and $b_k$. We found that a fixed $S$, $N1$ and $N2$ duration of 100 ms gave the highest performance. The weighting function used in (3) was the concatenation of three Hamming windows shown in Fig. 4 and the length of the maximum filter in (8) was set to $W = 30$ ms.

The input for the GMMs was a 14-component vector containing the estimated sibilant power spectrum from 1.5 kHz to 8 kHz every 500 Hz. The GMMs for sibilant and non-sibilant speech respectively used 14 and 28 full-covariance mixtures and were trained on the training subset of TIMIT. Sibilant phones and phones that sometimes include sibilant-like characteristics, such as stop consonants and non-sibilant fricatives, were excluded when training the non-sibilant GMM. To avoid problems caused by transcription alignment errors, phone transitions were omitted from the training. The SNR used for training was 0 dB in order to make the algorithm robust to noise, as it represents the lowest SNR at which sibilants/non-sibilants discrimination is practicable.

4. RESULTS

In this section, the performance of the proposed sibilant speech detector is evaluated. The results were calculated using the test set from the TIMIT database, which contains a total of 168 speakers and 1344 utterances. For evaluation purposes all non-sibilant phones were taken into account including stops and non-sibilant fricatives previously excluded for the training. Every time-frame was evaluated, without the removal of phone transitions.

White Gaussian noise was added to the speech files to generate the noisy test signals. The measurement of SNR used ITU-T P.56 [8, 9] for the speech level and unweighted power for the noise. Because of the noise-like nature of sibilant phones at high frequencies, we note that it is more difficult to detect sibilants in white noise than in other typical stationary noise sources where low frequencies often dominate.

The results obtained for $-5$ dB, 0 dB, 5 dB and 10 dB SNR as well as for clean speech are shown in Fig. 5. The detection error trade-off (DET) curves [10] in Fig. 5 shows the miss probability, $P_{miss}$, versus the false alarm probability, $P_{fa}$, as the likelihood ratio threshold is varied between 0.05 and 19.0. The equal error rates, where $P_{miss} = P_{fa}$, are listed in Table 1 and we see that at 0 dB SNR the equal error rate is 16.5% meaning that 83.5% of frames are correctly classified.

The circle on each line in Fig. 5 corresponds to a like-
llhood ratio threshold of unity corresponding to an estimated sibilant probability of 0.5. The values of $P_{\text{miss}}$ and $P_{\text{fa}}$ when using this threshold are listed in Table 2. We see that at SNR better than 10 dB, $P_{\text{miss}}$ is approximately constant at 11%. Manual inspection of these missed sibilant frames indicates that most of them correspond either to sibilant boundaries or to phones with very low energy. Below 10 dB SNR, the miss rate increases as sibilant phones are increasingly masked by the noise. In contrast, the false alarm rate, $P_{\text{fa}}$, remains below 17% at all SNR levels and declines at high noise levels when few sibilants are detected.

The sibilant detector described in this paper classifies each frame individually; it is possible that its classification accuracy could be further improved by applying temporal constraints to the classification decisions.

5. CONCLUSIONS

In this paper we have presented a sibilant detection algorithm robust to high levels of white Gaussian noise. The algorithm comprises a sibilant mean power estimation stage, which is based on a maximum likelihood approach, followed by a classification stage in which the likelihood ratio of two GMMs, one for sibilant speech and one for non-sibilant speech, is used. The algorithm has been evaluated on the TIMIT test set over a range of noise levels and consistently achieved over 80% classification accuracy for positive SNRs.

6. References


