A FREQUENCY-DOMAIN SILENCE NOISE MODEL

Guillaume Lathoud [a,b]
Mathew Magimai.-Doss [a,b]
Bertrand Mesot [a,b]

IDIAP–RR 05-13

APRIL 2005

a   IDIAP Research Institute, CH-1920 Martigny, Switzerland
b   Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland

A Frequency-Domain Silence Noise Model

Guillaume Lathoud   Mathew Magimai.-Doss   Bertrand Mesot

Abstract. This paper proposes a simple, computationally efficient 2-mixture model approach to discrimination between speech and background noise. It is directly derived from observations on real data, and can be used in a fully unsupervised manner, with the EM algorithm. A first application to sector-based, joint audio source localization and detection, using multiple microphones, confirms that the model can provide major enhancement. A second application to the single channel speech recognition task in a noisy environment yields major improvement on stationary noise and promising results on non-stationary noise.

1   Introduction

Robustness to various noise conditions is a key feature for speech processing algorithms to be turned into versatile, real-world applications. Most often, two exclusive directions are followed: either enhance the speech signal itself by ideally filtering out the noise [1, 2, 3], or change the way acoustic features are extracted from the signal [4, 5]. This paper presents an intermediary approach that enhances the feature extraction process at a level as close as possible to the original signal : at the magnitude spectrogram level, i.e. in time-frequency plane (TF). It relies on a 2-mixture model and unsupervised EM fitting [6] on observed data.

The underlying motivation of this approach is to rely on the estimated posterior probability of observing activity at a given (time, frequency) point of the spectrogram, so that ultimately the magnitude spectrogram can be replaced by a "posteriorgram". In spirit, the proposed approach can be related to TRAP-TANDEM [7] and further developments [8], although the probabilistic modeling is made here at a much lower-level type of data: the magnitude spectrogram itself.

Enhancing the spectrogram itself, based on probabilistic assumptions [9] has received much attention recently [2, 3]. In order to build a probabilistic model, at least two distributions are needed: one for background noise, and one for speech. A very reasonable model for background noise on silent parts of the TF plane is a white Gaussian assumption for real and imaginary parts, which translates into assuming a Rayleigh probability density function (pdf) in magnitude domain [10]. However, modeling of the speech part is much more complicate as such an assumption does not hold anymore. Super-gaussian models such as the Laplace pdf may be needed [2] for a better fit on real data. Derivation of the magnitude pdf of speech is then difficult, and still subject to research [11].

On the contrary, this paper proposes to restrict the problem to modeling of large magnitudes of speech only, implicitly attempting to drop the noisy part of the TF plane. In spirit, our approach is close to the missing data approach [12]. The main difference is that instead of a binary mask and the concepts of reliability and unreliability, the approach proposed here relies on the concepts of noise and activity, and aims at producing a "posteriorgram", which can be seen as a soft mask.

Intuitively, the main idea is that low speech magnitudes cannot be distinguished from background noise, being intrisically regions of low Signal-to-Noise Ratio (SNR). We therefore define activity in a *negative* way: activity is defined as what is *not* background noise[1]. The well-justified background noise Rayleigh model is therefore completed with an ad-hoc pdf for activity, that models "large" magnitudes only. "Large" is defined w.r.t. the Rayleigh model itself, and the complete modeling process is fully unsupervised. Two applications are considered: multimicrophone-based, enhanced sector-based speaker detection and localization, building on [14], and single channel noise-robust ASR. Both share the same generative model for the observed magnitude in TF plane. In the localization case, this model permits to detect and discard parts of the TF plane where the spatial point source model assumption does not hold. In the ASR case, the magnitude spectrogram is filtered at a reasonable cost, so that only speech that can be distinguished from background noise is retained.

The rest of this paper is organized as follows: Sect. 3 introduces the probabilistic model, motivating it with observations on real data. Sect. 4 reports detection/localization results on real meeting room data, and Sect. 5 reports ASR result on noisy telephone speech. Finally, Sect. 6 concludes.

2   Notations

Classical frame-based Fourier analysis is used. Both time t and frequency f are discretized into samples and N frequency bins (narrowbands), respectively.
- x_t is the captured signal,
- y_t is the preemphasized signal (in practice $y_t = x_t - 0.97 \cdot x_{t-1}$),
- $\mathcal{F}^y_{f,t}$ is the Digital Fourier Transform (DFT) of a windowed signal $[y_{t-N+1} \ldots y_t]^{\mathrm{T}}$ (using Hamming window),

---

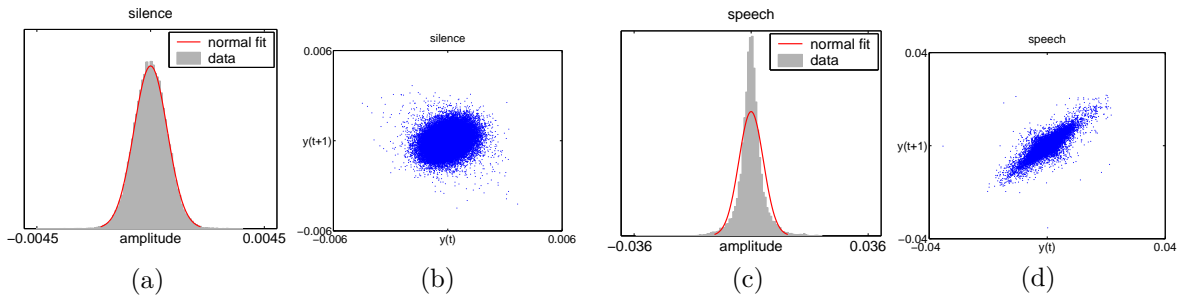[1]The authors are largely indebted to [13] for the "negative" and "structural" way of thinking.

Figure 1: Observations on real meeting room data [15] (pre-emphasized waveform $y(t)$). (a),(c): histograms, (b),(d): phase plots.

- $[\cdot]^{\mathrm{T}}$ denotes the transposition operator,
- $m_{f,t} = |\mathcal{F}^y_{f,t}|$ is the magnitude in TF plane,
- y and m designate realizations of random variables Y and M.

3 Proposed 2-Mixture Generative Model

In this section, the commonly used Rayleigh silence model is justified on real data, and completed with an ad-hoc "activity" model. The main difference with existing, related models such as in [9, 2, 3], is that we do not address the complete probabilistic modeling of speech activity, but limit ourselves to large magnitudes only.

3.1 Observations on Real Waveforms

Simple observations on silence periods of a pre-emphasized waveform $y(t)$ and its covariance matrix, as partially illustrated by Figs. 1a and 1b, show that modeling $\{Y_t\}$ as a i.i.d, zero-centered Gaussian processes is very reasonable. In particular, we computed the correlation matrix of several pre-emphasized silence signals $[y_1 \ldots y_N]^{\mathrm{T}}$ and observed that the correlation coefficients on the diagonal are at least one order of magnitude larger than others. Correlation and independence are equivalent for Gaussian random variables.

Formally, the time-domain model for silence can be written in the following manner:

$$\begin{cases} X_t & = & X_{t-1} + Y_t \\ Y_t & \sim & \mathcal{N}(0, \sigma_{\mathrm{sil}}) \\ \{Y_t\} & \text{i.i.d.} \end{cases} \tag{1}$$

where σ_{sil} is a parameter that does not depend on time t. Under such assumption, the real and imaginary part of the DFT are independent Gaussian distributed variables, as shown in Annex A. Note that this derivation is exact and does not rely on asymptotical considerations such as the central limit theorem. Thus, the magnitude $M_{f,t}$ has a Rayleigh pdf [10]. This type of assumption is used in several existing works [2, 11].

On the other hand, speech waveforms are clearly *not* Gaussian distributed, and *not* i.i.d., as shown by Fig. 1c and 1d. As mentioned previously, finding a fully-justified pdf for speech magnitude is still an open research subject. Hence, in Sect. 3.2 we propose to model large magnitudes of speech only.

3.2 Proposed Mixture Model

The proposed pdf for M is:

$$f(m) \stackrel{\text{def}}{=} P_{\mathrm{I}} \cdot f_{\mathrm{I}}(m) + P_{\mathrm{A}} \cdot f_{\mathrm{A}}(m), \tag{2}$$

where P_I and P_A are the priors for "silence" and "activity", respectively. From Eq 1 and Annex A, f_I is the Rayleigh pdf:

$$f_\mathrm{I}(m) \quad \overset{\mathrm{def}}{=} \quad \frac{m}{\sigma_\mathrm{I}^2} \; e^{-\frac{m^2}{2\sigma_\mathrm{I}^2}}, \tag{3}$$

and f_A is a pdf that models magnitudes $m > \delta_\mathrm{A}$, where δ_A is a threshold defined w.r.t. f_I. As a starting point, in this paper we use $\delta_\mathrm{A} = \sigma_\mathrm{I}$, which is the mode of the Rayleigh pdf. The reasoning is that values below the mode of the Rayleigh f_I can safely be assumed to be noise.

Moreover, we constrain f_A to fulfill two practical constraints. First, the derivative $f'_\mathrm{A}(m)$ of the chosen "activity" pdf should not be zero when m is just above δ_A, overwise the threshold δ_A will loose its meaning, as it may be set arbitrarily low. Second, the decay of $f_\mathrm{A}(m)$ when m tends towards infinity should be lower than the decay of the Rayleigh, in order to make sure that f_A will capture data with large magnitudes, and not f_I. A pdf that fulfills the two criterions above is a "shifted Erlang" pdf with h=2 (the Erlang pdf belongs to the Gamma family [10]):

$$f_\mathrm{A}(m) \quad \overset{\mathrm{def}}{=} \quad \mathbf{1}_{m>\delta_\mathrm{A}} \cdot \lambda_\mathrm{A}^2 \cdot (m - \delta_\mathrm{A}) \cdot e^{-\lambda_\mathrm{A}(m-\delta_\mathrm{A})}, \tag{4}$$

where $\mathbf{1}_{m>\delta_\mathrm{A}}$ is equal to 1 if $m > \delta_\mathrm{A}$, and zero otherwise. Note the implicit stationarity assumption: the 4 parameters $\Lambda = \{P_\mathrm{I}, \sigma_\mathrm{I}, P_\mathrm{A}, \lambda_\mathrm{A}\}$ are assumed independent of t. Independence of f is also assumed ; it is justified by the pre-emphasis, which whitens the spectrum.

EM training of Λ [6]: Both "E" and "M" steps involve simple mathematical expressions. The "M" step can be implemented by updating σ_I first, and then using the data above $\delta_\mathrm{A} = \sigma_\mathrm{I}$ to update λ_A. The cost can be further reduced by an histogram-based implementation.

Application: given an observed magnitude value $m_{f,t}$, and trained parameters Λ, the posterior probability of the activity is estimated as:

$$P\left(\mathrm{act}\,|m_{f,t},\ \Lambda\right) = \frac{P_\mathrm{A} \cdot f_\mathrm{A}(m_{f,t})}{P_\mathrm{I} \cdot f_\mathrm{I}(m_{f,t}) + P_\mathrm{A} \cdot f_\mathrm{A}(m_{f,t})}. \tag{5}$$

4 Application to Joint Source Detection and Localization

This section presents a joint detection/localization application of the mixture model presented in Sec. 3, that attempts to discard parts of the TF plane where the underlying spatial point-source model does not apply.

4.1 Sector-Based Beamforming

This section briefly reminds the sector-based joint detection and localization approach described in full details in [14]. The space around a microphone array is divided into a finite number of sectors. Using phase information only, in a given sector, this approach detects whether the sector contains at least one active source or zero. Let Q be the number of microphones, and P be number of pairs of microphones (i_p, j_p), where $1 \le i_p < j_p \le Q$. When all possible pairs are used, $P = Q(Q-1)/2$. Based on [14], the average "phase-only" beamforming over a given volume of space is defined as:

$$E_{f,t}^{\mathrm{ds}} \quad \overset{\mathrm{def}}{=} \quad \frac{1}{P} \sum_{p=1}^{P} E_{f,t,p}^{\mathrm{ds}} \quad \text{where}: \tag{6}$$

$$E_{f,t,p}^{\mathrm{ds}} \quad \overset{\mathrm{def}}{=} \quad 1 + A_{f,p}(\mathbf{v}) \cdot \cos\left(\hat{\theta}_{f,t,p} - B_{f,p}(\mathbf{v})\right), \tag{7}$$

where $A_{f,p}$ and $B_{f,p}$ are fixed, real-valued parameters, derived from the average theoretical cross-correlation matrix over volume $\mathbf{v}$, and $\hat{\theta}_{f,t,p}$ is the measured phase difference at frequency f, between

microphones i_p and j_p. $A_{f,p}$ and $B_{f,p}$ are independent of the measured data, and are computed only once for each geometrical configuration of the array and the sectors.

Within each frequency bin f, Eq. 6 is used to determine the "most active" sector, i.e. the sector that has maximum $E_{f,t}^{\mathrm{ds}}$. The activity of each sector is then defined as the total number of frequency bins where it is dominant, in a given time frame. Although [16, 14] have shown very good overall performance of this approach, magnitude is not used, thus leading to random decisions and suboptimal results on silent parts of the TF plane.

4.2 Modified Sector-Based Beamforming

Global frame energy is not a good activity detector for source localization [17]. We thus attempt here to incorporate magnitude information at the frequency bin level, into Eq. 6. Two ways are proposed, both define magnitude-based weight functions $w(m) \geq 0$ in order to activate or deactivate each microphone pair p. Eq. 6 is replaced with:

$$E_{f,t}^{\mathrm{dsw}} \quad \stackrel{\mathrm{def}}{=} \quad \frac{1}{P} \sum_{p=1}^{P} w\left(m_{f,t}^{(i_p)}\right) \cdot w\left(m_{f,t}^{(j_p)}\right) \cdot E_{f,t,p}^{\mathrm{ds}} \tag{8}$$

The first way uses magnitude itself: $w_{\mathrm{MAG}}(m) = m$.

The second way relies instead on the mixture model proposed in Sect. 3.2: $w_{\mathrm{POST}}(m) = P(act|m, \Lambda)$. Note that in this case, any time that one microphone has its measured magnitude m less than the automatic threshold δ_{A}, all pairs using this microphone will be zeroed in Eq. 8.

The main difference between these two ways is that the posterior-based weighting function w_{POST} implicitly discards any microphone-specific gain information, thus removing the need for precise microphone calibration. Obtaining with w_{POST} a detection/localization performance similar to w_{MAG} would therefore be an interesting achievement, with practical implications: e.g. robustness to long-term drift of the calibrated gain, user-friendliness of a portable meeting capture system.

4.3 Experimental Results

Five real 16kHz audio sequences were taken from a meeting room audio-visual corpus available online [15], recorded with a horizontal circular 8-mic array (10 cm radius) set on a table. Complete data and description can be found at `http://mmm.idiap.ch/Lathoud/05-ICASSP` Total duration is 1 hour 13 minutes. 1 hour of the data has either 2 or 3 concurrent loudspeakers playing controlled, synthetic speech. In addition, 13 min of the data contains speech from up to 3 real humans, speaking mostly concurrently.

In a previous work using the same data [16], a threshold on "activity" (total number of active frequency bins for a given sector, in a given time frame) was set on a reduced development set and kept fixed on the rest of the data (test set). However, we noted that the distributions of "activity" on *truly* active/inactive (sector, time frame) pairs can vary noticeably between recordings, and, for some methods, between speech and silence periods.

Therefore, in this paper the evaluation is done in a fully unsupervised manner. This removes the need for a development set and a fixed threshold. Each recording is processed independently, with the following steps:

1. Compute activity, i.e. the total number of active frequency bins, for each (sector, time frame).

2. Fit two Rice distributions [18] on all activity values, irrespective of the sector and time frame. This is done using the EM algorithm [6]. The M step is approximated with the simple and efficient moment method proposed in [18]. After convergence, the Rice distribution with lowest (resp. highest) average power is interpreted as "silence" (resp. "speech").

3. Based on the desired working point (e.g. False Alarm Rate of 0.5%) and the cumulative function of each Rice distribution, select a threshold.

Case	0 to 3 loudspeakers				1 human	0 to 3 humans
Sequence	#1	#2	#3	Average	#4	#5
Target	0.5, 0	0.5, 0	0.5, 0	0.5, 0	0.5, ≈32.6	0.5, ≈32.6
Baseline	1.18, 7.53	1.62, 9.26	1.99, 5.35	1.60, 7.38	1.94, 33.7	2.98, 44.1
w_{MAG}	1.42, 8.03	1.36, 9.83	1.08, 9.42	1.29, 9.09	1.84, 32.1	3.54, 40.5
w_{MAG}, uncal.	1.66, 15.0	1.45, 19.5	1.44, 15.1	1.52, 16.5	1.36, 36.3	3.27, 46.5
w_{POST}	1.86, **1.79**	1.76, **6.76**	2.04, **0.85**	1.89, **3.13**	1.86, 32.9	4.42, 42.1
w_{POST}, uncal.	1.86, **1.79**	1.76, **6.76**	2.04, **0.85**	1.89, **3.13**	1.86, 32.9	4.42, 42.1

(a) Overall (FAR, FRR) in percentages.

Case	2 loudspeakers				2 humans
Sequence	#1	#2	#3	Average	#5
Target	2.0	2.0	2.0	2.0	≈1.35
Baseline	1.99	**1.95**	1.99	1.98	1.22
w_{MAG}	1.99	**1.95**	1.98	1.97	1.28
w_{MAG}, uncal.	1.92	1.88	1.92	1.91	1.17
w_{POST}	**2.00**	**1.95**	**2.00**	**1.98**	1.23
w_{POST}, uncal.	**2.00**	**1.95**	**2.00**	**1.98**	1.23

(b) $\overline{N_c}$, 2-source case

Case	3 loudspeakers				3 humans
Sequence	#1	#2	#3	Average	#5
Target	3.0	3.0	3.0	3.0	≈2.02
Baseline	2.72	2.64	2.78	2.71	1.51
w_{MAG}	2.70	2.62	2.62	2.65	1.63
w_{MAG}, uncal.	2.46	2.25	2.41	2.37	1.44
w_{POST}	**2.93**	**2.75**	**2.97**	**2.88**	1.62
w_{POST}, uncal.	**2.93**	**2.75**	**2.97**	**2.88**	1.62

(c) $\overline{N_c}$, 3-source case

Table 1: Joint detection/localization results. "Uncal." denotes uncalibrated microphones (random gain, uniform over [-12,+12] dB). FAR and FRR are estimated over *all* (sector, time frame) pairs. $\overline{N_c}$ is the average number of active sources that are *simultaneously* detected & correctly located. **Targets for humans are approximate** because segments annotated as "speech" contains short silences.

Tab. 1 reports the results. 16 ms frame shift was used, with 32 ms frame length. Note that the loudspeaker result only can be used for numerical comparisons between methods (exact target), while results on human data are more of a sanity check (approximate target).

Overall, both types of weighting functions achieve a significant improvement over the baseline, with decent results on human data. The w_{POST} approach achieves the best results in the 2- and 3-loudspeaker cases, and its advantage over w_{MAG} in the case of uncalibrated microphones is clear.

5 Application to Noise-Robust ASR

This section presents a noise-robust ASR application of the mixture model proposed in Sec. 3, that attempts to enhance the MFCC feature extraction process for enhanced robustness to noise. The context is HMM/GMM speech recognition.

5.1 Baseline MFCC Extraction

12.5 ms frame shift is used, with 32 ms frame length. At each time frame t, MFCC extraction is implemented as follows:

- **Step 1:** The magnitude spectrum $[m_{1,t} \ldots m_{N,t}]^{\mathrm{T}}$ is estimated, as explained in Sect. 2.

- **Step 2:** Mel-filterbanks, log compression and Digital Cosine Transform (DCT) are applied to $[m_{1,t} \ldots m_{N,t}]^{\mathrm{T}}$, yielding cepstral coefficients $c_t^0, \ldots, c_t^{12}$.

- **Step 3:** Mean-removed cepstral coefficients, along with their deltas and delta-deltas (39 dimensions), are fed into the HMM/GMM system for training it or testing it.

5.2 Modified MFCC Extraction

Modification of Step 1: we simply propose to fit the mixture model using the EM algorithm, as presented in Sec. 3, and use the posteriors of activity (or silence) to filter or replace the magnitude spectrogram. Three methods are proposed:

- **"POSTFILT":** In words, all information below the automatic threshold δ_{A} is dropped, and spectral peaks are emphasized. Formally:

$$m_{f,t}^{\mathrm{POST}} \quad \stackrel{\mathrm{def}}{=} \quad 1 + \left(\frac{m_{f,t}}{\delta_{\mathrm{A}}} - 1 \right) \cdot P\left(\mathrm{act} \,|\, m_{f,t},\, \Lambda\right) \qquad (9)$$

  When $m_{f,t} \leq \delta_{\mathrm{A}}$, $m_{f,t}^{\mathrm{POST}} = 1$, otherwise $m_{f,t}^{\mathrm{POST}} > 1$. The purpose of the division by δ_{A} is to normalize out the possible variations in microphone gain, from one file to another.

- **"POWERFILT":** It is very similar to "POSTFILT", but spectral peak enhancement is achieved in a different fashion:

$$m_{f,t}^{\mathrm{POWER}} \quad \stackrel{\mathrm{def}}{=} \quad \left(\frac{m_{f,t}}{\delta_{\mathrm{A}}} \right)^{P(\mathrm{act}|m_{f,t},\, \Lambda)} \qquad (10)$$

- **"PSIL":** the original intent was to use $[P\left(\mathrm{sil} \,|\, m_{f,t}, \Lambda\right)]^{-1}$ as a feature to replace magnitude, given that the ratio $\frac{f_{\mathrm{A}}(m)}{f_{\mathrm{I}}(m)}$ increases exponentially when m is large (spectral peaks), as discussed in Sec. 3.2. However, the dynamic range involved is very often too large for the numerical limits of a standard computer, so we had to compress it using a log function:

$$m_{f,t}^{\mathrm{PSIL}} \quad \stackrel{\mathrm{def}}{=} \quad -\log\left[\min\left(1 - \epsilon, P\left(\mathrm{sil} \,|\, m_{f,t},\, \Lambda\right)\right)\right], \qquad (11)$$

  $\epsilon > 0$ is a constant that should be small, otherwise too much information is lost. In experiments, we did a minor tuning on ϵ, trying $0.01, 0.05$ and 0.1. $\epsilon = 0.05$ gave the best results.

Condition:	clean	Factory Noise			Lynx Noise		
SNR (dB)	∞	0	6	12	0	6	12
Baseline	93.5	12.7	59.7	84.9	50.6	81.4	90.0
CJ-RASTA-PLP[†]	90.2	43.2	74.8	**86.7**	65.1	82.1	88.6
PAC[†]	87.8	**50.8**	**75.5**	85.9	64.5	79.4	86.0
POSTFILT	91.9	36.9	69.3	83.8	74.4	86.1	89.8
POSTFILT (block)	91.7	38.5	69.0	85.1	74.1	86.3	90.1
POWERFILT	91.7	38.1	68.8	83.0	74.7	85.8	89.7
POWERFILT (block)	91.1	43.1	70.7	84.7	**75.4**	**86.5**	90.1
PSIL	**92.9**	41.9	71.3	84.3	72.4	84.4	**90.3**
PSIL (block)	91.7	47.9	71.7	84.4	73.5	85.1	89.7

Table 2: Word Accuracy on OGI Numbers including several existing approaches († denotes results given in [21]), and the proposed approaches "POSTFILT", "POWFILT" and "PSIL". "block" denotes block-wise processing (0.25 s, i.e. 20 frames). **Bold face** indicates the best non-baseline result in each column.

5.3   Global Activity Feature

The modified MFCCs were found to be noise-robust in preliminary experiments. In addition, a "global activity" feature $\Omega_t \in [0, 1]$ is defined as:

$$\Omega_t \quad \stackrel{\text{def}}{=} \quad \frac{1}{N} \sum_{f=1}^{N} P\left(\text{act} \,|m_{f,t}\right). \tag{12}$$

This feature is a probabilistic estimate of the bandwidth occupied by activity, at a given time frame t. After utterance-level mean and standard deviation normalization, we found that replacing c_t^0 with Ω_t allows for a major additional improvement in performance, especially in noisy conditions. A possible interpretation is that after utterance-level normalization, Ω_t is a reliable speech presence indicator.

5.4   Experimental Results

All experiments reported here with the methods presented in Sec. 5.2 use $\left[\Omega_t, c_t^1, \ldots, c_t^{12}\right]^{\mathrm{T}}$, computed on the filtered magnitude spectrum, along with their deltas and delta-deltas.

The baseline result is using, on the original magnitude spectrum, $\left[c_t^0, c_t^1, \ldots, c_t^{12}\right]$, their deltas and delta-deltas.

The OGI Numbers database [19] is used for connected word recognition, with respectively 3233 and 1206 utterances in the training and test sets. Only "clean" conditions are used for training. For testing, in addition to the original "clean" conditions, the non-stationary "Factory" noise and the stationary "Lynx" noise from the NoiseX 92 database [20] were added at three levels: 0, 6 and 12 dB.

Since the proposed model is inherently stationary, higher enhancement is expected on "Lynx" than on "Factory". We thus ran the experiments twice: once offline, and once processing each file in a block-wise fashion. Results are reported in Tab. 2, along with those of two state-of-the-art noise-robust approaches. Overall, all three proposed methods behave similarly, obtaining the best results on all "Lynx" conditions. Moreover, on "clean" conditions, they achieve significantly higher performance than CJ-RASTA-PLP and PAC.

As expected, there is room for improvement in non-stationary conditions, although: (1) results are encouraging, with significant improvement over the MFCC baseline, (2) the blockwise processing hints at strong potential for further improvement.

6   Conclusion

A simple, inexpensive and effective 2-mixture generative model was proposed to discriminate between noise and speech in the TF plane. A key point is that the speech mixture component only models large magnitudes. The 2-mixture model is trained on observed data in a fully unsupervised manner, using the EM algorithm. Two applications are given that validate the model, showing major improvements. In both cases, the key idea is to use the posterior probability of activity in the TF plane. On the audio source detection and localization side, close-to-perfect detection of up to 3 concurrent sources was obtained on real data. Avenues for future research include investigating alternate weighting strategies and possible extension to fine 3-D localization of multiple sources. On the noise-robust ASR side, a major improvement was obtained over existing approaches on both clean and stationary noise conditions. In non-stationary noise conditions, ASR results are encouraging, and preliminary "block-wise" results showed strong potential for improvement. Directions for future work include large vocabulary conversational speech recognition.

7   Acknowledgements

Annex A

In this Section we derive the Rayleigh magnitude-domain silence model of $|\mathcal{F}_f^Y|$ (Eq. 3) from the white Gaussian time-domain silence model of Y_t (Eq. 1).

First, let us recall a result originally shown by Rice in 1944 (for a demonstration see [10, pp. 296-297]). Rice showed that given two zero-mean Gaussian, uncorrelated random variables A and B with same standard deviation σ, and $R \overset{\text{def}}{=} |A + jB|$, the R variable has a Rayleigh pdf:

$$f_R(r) = \frac{r}{\sigma} e^{-\frac{r^2}{2\sigma^2}} \quad \text{for} \quad r > 0 \tag{13}$$

Let us now define $Y_{1:N}$, as in Eq. 1, as a vector of N uncorrelated[2] zero-mean Gaussian random variable $Y_{1:N} \overset{\text{def}}{=} [Y_1 \ldots Y_N]^{\mathrm{T}}$. The Discrete Fourier Transform (DFT) of Y is $\mathcal{F}_{1:N}^Y = \left[\mathcal{F}_1^Y \cdots \mathcal{F}_N^Y\right]^T$ where for a given $f = 1 \ldots N$:

$$\mathcal{F}_f^Y \overset{\text{def}}{=} \sum_{n=1}^{N} Y_n e^{-2\pi (f-1)\frac{n-1}{N}} \tag{14}$$

Let $A_f = \mathcal{R}e\left(\mathcal{F}_f^Y\right)$ and $B_f = \mathcal{I}m\left(\mathcal{F}_f^Y\right)$. In other terms:

$$\begin{cases} A_f = \sum_{n=1}^{N} Y_n \cos\left(-2\pi (f-1)\frac{n-1}{N}\right) \\ B_f = \sum_{n=1}^{N} Y_n \sin\left(-2\pi (f-1)\frac{n-1}{N}\right) \end{cases} \tag{15}$$

For $f = 1$ we have $A_1 = \sum_{n=1}^{N} Y_n = 0$ and $B_1 = 0$.

For $f > 1$: the random variable A_f (resp. B_f) is a weighted sum of zero-mean, single Gaussian random variables, therefore [22, p. 99] it is also a zero-mean, single Gaussian random variable with variance:

$$\begin{cases} \sigma_{A_f}^2 = \sigma^2 \sum_{n=1}^{N} \cos^2\left(2\pi (f-1)\frac{n-1}{N}\right) \\ \sigma_{B_f}^2 = \sigma^2 \sum_{n=1}^{N} \sin^2\left(2\pi (f-1)\frac{n-1}{N}\right). \end{cases} \tag{16}$$

---

[2]Uncorrelation and independence are equivalent for Gaussian random variables.

Given that $\cos^2 t = \frac{1}{2}(1 + \cos 2t)$ and $\sin^2 t = \frac{1}{2}(1 - \cos 2t)$ we can write:

$$\begin{cases} \sigma_{A_f}^2 = \frac{\sigma^2}{2} \left(N + \sum_{n=1}^{N} \cos \left(4\pi \left(f - 1 \right) \frac{n-1}{N} \right) \right) \\ \sigma_{B_f}^2 = \frac{\sigma^2}{2} \left(N - \sum_{n=1}^{N} \cos \left(4\pi \left(f - 1 \right) \frac{n-1}{N} \right) \right). \end{cases} \tag{17}$$

Let us now write the complex domain sum:

$$\sum_{n=1}^{N} e^{j4\pi(f-1)\frac{n-1}{N}} = \sum_{n=0}^{N-1} \alpha^n = \frac{1 - \alpha^N}{1 - \alpha} = 0, \tag{18}$$

because $\alpha^N = 1$, where $\alpha = e^{j4\pi\frac{f-1}{N}}$. (Since $1 < f \le N$, $\alpha \ne 1$ and all terms in Eq. 18 are defined.) From Eqs. 17 and 18 we conclude that:

$$\sigma_{A_f} = \sigma_{B_f} = \sigma\sqrt{\frac{N}{2}}. \tag{19}$$

As for the cross-correlation $\sigma_{A_f B_f} \overset{\text{def}}{=} \mathbf{E}\{A_f B_f\}$, it is a weighted sum of $\mathbf{E}\{Y_n Y_p\}$ which are all zero because of the uncorrelation hypothesis, therefore $\sigma_{A_f B_f} = 0$.

To conclude, we have shown that the random variables A_f and B_f are zero-mean, uncorrelated single Gaussian random variables of same variance, therefore the result of Rice applies:

For $f > 1$, $\left| \mathcal{F}_f^Y \right|$ has a Rayleigh pdf of parameter $\sigma\sqrt{\frac{N}{2}}$.

References

[1] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001, ch. 2, pp. 19–38.

[2] R. Martin and C. Breithaupt, "Speech enhancement in the dft domain using Laplacian speech priors," in *Proc. of IWAENC 2003*, 2003.

[3] R. Gemello, F. Mana, and R. D. Mori, "A modified Ephraim-Malah noise suppression rule for automatic speech recognition," in *Proc. of ICASSP 2004*, 2004.

[4] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proceedings of ICASSP'92*, 1992.

[5] S.Ikbal, H.Misra, and H.Bourlard, "Phase autocorrelation (PAC) derived robust speech features," in *Proceedings of ICASSP 2003*, 2003.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," vol. 39, pp. 1–38, 1977.

[7] H. Hermansky, "TRAP-TANDEM: Data-driven extraction of temporal features from speech," in *Proceedings of ASRU 2003*, 2003.

[8] H. Bourlard, S. Bengio, M. M. Doss, Q. Zhu, B. Mesot, and N. Morgan, "Towards using hierarchical posteriors for flexible automatic speech recognition systems," in *Proceedings of the DARPA EARS RT04 Workshop*, 2004.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," in *Proceedings of ICASSP 1984*, 1984.

[10] C. M. Grinstead and J. L. Snell, *Introduction to Probability*. American Mathematical Society, 1997.

[11] B. Chen and P. Loizou, "Speech enhancement using a MMSE short time spectral amplitude estimator with Laplacian speech modeling," in *Proc. of ICASSP 2005*.

[12] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[13] A. Korzybski, *Science and Sanity*. The International Non-Aristotelian Library Publishing Company, 1933.

[14] G. Lathoud, J. Bourgeois, and J. Freudenberger, "Sector-based detection for hands-free speech enhancement in cars," IDIAP-RR-04- 67, 2004.

[15] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking," in *Proceedings of the 2004 MLMI Workshop, S. Bengio and H. Bourlard Eds, Springer Verlag*, 2005.

[16] G. Lathoud and M. Magimai.-Doss, "A Sector-Based, Frequency-Domain Approach to Detection and Localization of Multiple Speakers," in *Proc. of ICASSP'05*, 2005.

[17] G. Lathoud and I. McCowan, "A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays," in *Proc. SAPA 2004*, Oct. 2004.

[18] L. Greenstein, D. Michelson, and V. Erceg, "Moment-method estimation of the ricean K-factor," *IEEE Communications Letters*, vol. 3, no. 6, pp. 175–176, june 1999.

[19] R. A. Cole, M. Fanty, M. Noel, and T. Lander, "Telephone speech corpus development at CSLU," 1994.

[20] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," DRA Speech Research Unit, Malvern, England, Tech. Rep., 1992.

[21] S. Ikbal, "Nonlinear feature transformations for noise robust speech recognition," Ph.D. dissertation, EPFL, Switzerland, 2004.

[22] J.-P. Delmas, *Introduction aux Probabilités*. Ellipses Marketing, 1993.