IDIAP RESEARCH REPORT

Unsupervised Spectral Subtraction for Noise-Robust ASR

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

IDIAP–RR 05-42

July 2005

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

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-42

Unsupervised Spectral Subtraction for Noise-Robust ASR

Guillaume Lathoud   Mathew Magimai.-Doss   Bertrand Mesot
Hervé Bourlard

July 2005

Abstract. This paper proposes a simple, computationally efficient 2-mixture model approach to discriminate between speech and background noise at the magnitude spectrogram level. It is directly derived from observations on real data, and can be used in a fully unsupervised manner, with the EM algorithm. In this paper, the 2-mixture model is used in an "Unsupervised Spectral Subtraction" scheme that can be applied as a pre-processing step for any acoustic feature extraction scheme, such as MFCCs or PLP. The goal is to improve noise-robustness of the acoustic features. Experimental results on both OGI Numbers 95 and Aurora 2 tasks yielded a major improvement on all noise conditions, while retaining a similar performance on clean conditions.

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 is related to TRAP-TANDEM [7] and further developments [8], although the probabilistic modeling is made here at a much lower level: the magnitude spectrogram itself.

Enhancing the spectrogram itself, based on probabilistic assumptions [9], led to applications to noise-robust ASR [10, 11], and 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 [12]. However, modeling of the speech part is much more complicate as such an assumption does not hold anymore. Supergaussian 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 [13].

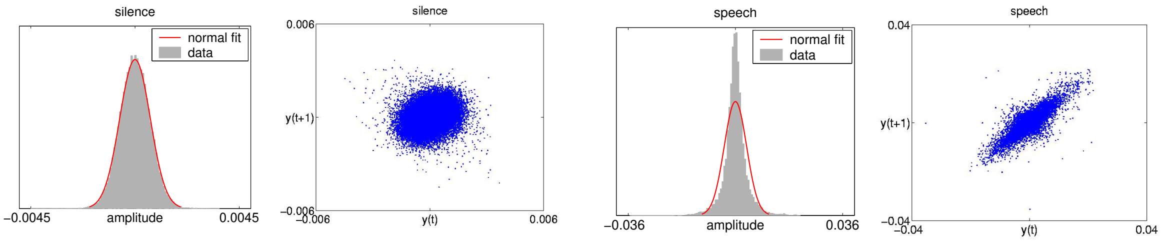
On the contrary, this paper proposes to restrict the problem to the modeling of large magnitudes of speech only. Intuitively, the main idea is that low speech magnitudes cannot be distinguished from background noise, being intrinsically regions of low Signal-to-Noise Ratio (SNR). We therefore complete the well-justified background noise Rayleigh model 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. We apply this approach to enhance the noise-robustness of single channel ASR by removing noise from the magnitude spectrogram using spectral subtraction. The magnitude spectrogram is filtered at a small cost, so that only speech that can be distinguished from background noise is retained. No parameter tuning is involved, thus justifying the "unsupervised spectral subtraction" designation.

Note that the purpose of this paper is *not* to propose novel noise-robust ASR features, but rather a simple, generic approach that can be used as a pre-processing step to any acoustic feature extraction process (MFCCs, PLP, etc.). This type of approach is very much in the spirit of [11]. The present paper constitutes mainly an update of [14]. The main contributions are a simplification of the approach, given in Sect. 3.3 and a confirmation of the OGI Numbers 95 results by testing on the Aurora 2 task, given in Sect. 4.4.

The rest of this paper is organized as follows: Sect. 3 introduces the probabilistic model, motivating it with observations on real data. The proposed approach is contrasted to the closest existing approaches such as [10, 11]. Sect. 4 reports ASR result on noisy telephone speech: OGI Numbers 95 [15], as well as Aurora 2 [16]. Finally, Sect. 5 concludes.

2 Notations

Both time t and frequency f are discretized into samples and N frequency bins (narrowbands), respectively. y_t is the pre-emphasized signal, $\mathcal{F}^y_{f,t}$ is the Discrete Fourier Transform (DFT) of a windowed signal $[y_{t-N+1} \dots y_t]^{\mathrm{T}}$ (using Hamming window), $[\cdot]^{\mathrm{T}}$ 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.



(a) (b) (c) (d)

Figure 1: Observations on real meeting room data [17] (pre-emphasized waveform $y(t)$). (a),(c): histograms, (b),(d): phase plots.

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 process is very reasonable. 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 [12]. This type of assumption is used in several existing works [2, 13].

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 $f\,(m) \stackrel{\text{def}}{=} P_{\text{I}} \cdot f_{\text{I}}(m) + P_{\text{A}} \cdot f_{\text{A}}(m)$, where P_{I} and P_{A} are the priors for "silence" and "activity", respectively. Note that the model is independent of f and t.

f_{I} is the Rayleigh pdf:

$$f_{\text{I}}(m) \quad \stackrel{\text{def}}{=} \quad \frac{m}{\sigma_{\text{I}}^2} \; e^{-\frac{m^2}{2\sigma_{\text{I}}^2}}, \tag{1}$$

and f_{A} is a pdf that models magnitudes $m > \delta_{\text{A}}$, where δ_{A} is a threshold defined w.r.t. f_{I}. As a starting point, in this paper we use $\delta_{\text{A}} = \sigma_{\text{I}}$, 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 background noise.

Moreover, we constrain f_{A} to fulfill two practical constraints. First, the derivative $f'_{\text{A}}(m)$ of the chosen "activity" pdf should not be zero when m is just above δ_{A}, otherwise the threshold δ_{A} will loose its meaning, as it may be set to an arbitrarily low value. Second, the decay of $f_{\text{A}}(m)$ 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 [12]):

$$f_{\text{A}}(m) \quad \stackrel{\text{def}}{=} \quad \mathbf{1}_{m>\sigma_{\text{I}}} \cdot \lambda_{\text{A}}^2 \cdot (m - \sigma_{\text{I}}) \cdot e^{-\lambda_{\text{A}}(m-\sigma_{\text{I}})}, \tag{2}$$

where $\mathbf{1}_{m>\sigma_\mathrm{I}}$ is equal to 1 if $m > \sigma_\mathrm{I}$, 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 to be independent of t. Furthermore, 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. In the "E" step, posteriors can be estimated as follows:

$$P\left(\mathrm{sil}\,|m_{f,t},\ \Lambda\right) \quad = \quad \frac{P_\mathrm{I} \cdot f_\mathrm{I}(m_{f,t})}{P_\mathrm{I} \cdot f_\mathrm{I}(m_{f,t}) + P_\mathrm{A} \cdot f_\mathrm{A}(m_{f,t})}. \tag{3}$$

and $P\left(\mathrm{act}\,|m_{f,t},\ \Lambda\right) = 1 - P\left(\mathrm{sil}\,|m_{f,t},\ \Lambda\right)$.

In the "M" step, exact maximization of the likelihood is difficult: since parameters σ_I and λ_A are tied in a non-linear fashion (Eq. 2), we have not found a way to separate them analytically. From this constatation, we chose two options:

1. Numerical optimization of the exact likelihood of the observed data though simplex search in the $(\sigma_\mathrm{I}, \lambda_\mathrm{A})$ space (e.g. `fminsearch` in MATLAB). In this case, the M step requires several sub-iterations before reaching a local maximum. We refer to this option as "**ML**".

2. Analytical approximation through moment-method (a single step). First σ_I is updated using all data, then λ_A is updated using all data with values above the new σ_I value. This option is referred to as "**moment**". It is defined by the following update equations:

$$\hat{\sigma}_\mathrm{I} \quad = \quad \left[\frac{\sum_{f,t} m_{f,t}^2 \cdot P\left(\mathrm{sil}\,|m_{f,t},\ \Lambda\right)}{2 \sum_{f,t} P\left(\mathrm{sil}\,|m_{f,t},\ \Lambda\right)}\right]^{\frac{1}{2}}$$

$$\hat{\lambda}_\mathrm{A} \quad = \quad \frac{\sum_{m_{f,t} > \hat{\sigma}_\mathrm{I}} \left(m_{f,t} - \hat{\sigma}_\mathrm{I}\right)^{-1} \cdot P\left(\mathrm{act}\,|m_{f,t},\ \Lambda\right)}{\sum_{m_{f,t} > \hat{\sigma}_\mathrm{I}} P\left(\mathrm{act}\,|m_{f,t},\ \Lambda\right)}$$

Data representation: While in [14] a small-cost histogram-based implementation was used, in the present paper the cost is further reduced by opting for a direct representation with a reduced set of samples (no weights). In all results reported in this paper, only 100 representative samples are used: observed magnitude values are ordered, from which 100 samples are picked at regular intervals. This is equivalent to estimate percentile values at equal steps.

An example of fit on one file taken from the OGI Numbers 95 database [15] can be seen in Figs. 2a, 2b and 2c.

3.3 Unsupervised Spectral Subtraction (USS)

A 2-step approach is used:

1. EM fitting of the 2-mixture model, as described in Sect.3.2. Either the **ML** method or the **moment** approximation can be used.

2. Spectral subtraction using parameter σ_I as a floor:

$$m_{f,t}^{\mathrm{USS}} \quad \stackrel{\mathrm{def}}{=} \quad \max\left(1, \frac{m_{f,t}}{\sigma_\mathrm{I}}\right) \tag{4}$$

Note that the flooring to a non-zero value $(\max(1, \ldots))$ is necessary in the MFCC context. Indeed, leaving zero magnitude values after spectral subtraction would lead to undesirable dynamics in cepstral coefficients. An example of result of Eq. 4 is shown in Fig. 2d.

(a) Original spectrogram ($m_{f,t}$) (b) Fit (linear domain) (c) Fit (log domain) (d) Spectrogram after subtraction ($m^{\mathrm{USS}}_{f,t}$)

Figure 2: Example of fit of the 2-mixture model on noisy data taken from the OGI Numbers 95 database (Factory 0dB condition). All plots show magnitude data in the frequency domain. On spectrogram plots (a) and (d), the largest magnitudes are red, the smallest magnitudes are blue.

This approach can be compared to previous works. We can note common points "in spirit" with [11] (and to a lesser extent [10]): adaptation to non-stationary noises is possible in both [11] and our approach through block-wise processing, as illustrated by experiments reported in Sect. 4.3. Moreover, on the modeling side, parameters of the "noise" distribution (i.e. silence) are more important than those of the "speech" distribution (i.e. activity) in both approaches, as they are used to floor the magnitude values (max operator).

However, here the modeling is made directly at the magnitude spectrogram level, with a single model for all frequencies, while in [11] modeling was made after mel filterbank computation, and a model was defined for each critical band separately. Moreover, the approach proposed here is one-pass, fully unsupervised, without any "feedback" loop, without any tunable threshold, without histogram. In [11], multiple stages are involved, including histograms, band-specific parameters, short-term and long-term adaptation, a "feedback" loop, tunable parameters that have to be trained and (optionally) injection of an artificial white noise.

Finally, note that all posterior-based filtering approaches previously proposed in [14] yielded inferior ASR results, as compared to the simple spectral subtraction approach of Eq. 4. A consequence is that posterior computation for spectrogram filtering is not necessary anymore, only comparing magnitude values $m_{f,t}$ to the σ_{I} parameter is needed. The computational cost is therefore reduced.

4 Application to Noise-Robust ASR

This section presents an application of the 2-mixture model proposed in Sect. 3, that attempts to enhance the MFCC feature extraction process for greater robustness to noise. The context is HMM/GMM speech recognition.

4.1 Baseline MFCC Extraction

Overlapping time frames are used, for example 12.5 ms frame shift is used, and 32 ms frame length. At each time frame t, MFCC extraction is implemented as follows:

- **Step 1:** The magnitude spectrum $[m_{1,t} \dots m_{N,t}]^{\mathrm{T}}$ is estimated, as explained in Sect. 2.
- **Step 2:** Mel-filterbanks, log compression and Discrete Cosine Transform (DCT) are applied to $[m_{1,t} \dots m_{N,t}]^{\mathrm{T}}$, yielding cepstral coefficients $c_t^0, \dots, c_t^{12}$.
- **Step 3:** Mean- and variance-normalized cepstral coefficients, along with their deltas and delta-deltas (39 dimensions), are fed into the HMM/GMM system for training it or testing it. This is true for all experiments reported in this paper.

One difference with results initially reported in [14] is the use of variance normalization, which dramatically improves all results. Note that both mean and variance normalizations are applied at the utterance level, before delta and delta-delta estimation.

Condition: SNR (dB)	clean ∞	Factory Noise 0	6	12	Lynx Noise 0	6	12
Baseline	6.8	50.3	23.9	13.3	25.0	14.7	10.1
Proposed (ML)	**6.7**	47.9	25.0	13.5	21.5	13.2	9.2
Proposed (moment)	**6.7**	46.0	22.2	12.7	**19.8**	**12.1**	**8.9**
Proposed (moment, block)	7.8	**43.9**	**21.5**	**12.5**	19.9	12.5	9.4

Table 1: Word Error Rate (WER) on OGI Numbers 95. Both cepstral mean and variance normalizations are used in all cases. "block" denotes block-wise processing (0.25 s, i.e. 20 frames). **Bold face** indicates the best result in each column.

4.2 MFCC Extraction after Spectral Subtraction

Modification of Step 1: we simply propose to fit the mixture model using the EM algorithm, as presented in Sect. 3.2, and apply the unsupervised spectral subtraction as described in Sect. 3.3. MFCCs are subsequently extracted, exactly as in 4.1.

- Note 1: this simple spectral subtraction scheme yielded better results than any of the various posterior-based filtering approaches initially proposed in [14]. Therefore, only results obtained with spectral subtraction are reported in this paper.
- Note 2: cepstral coefficients $c_t^0, \ldots, c_t^{12}$ are used for recognition. It yielded better results than the "global activity feature" initially proposed in [14].

4.3 Experimental Results on OGI Numbers 95 Task

12.5 ms frame shift is used, and 32 ms frame length. The OGI Numbers 95 database [15] 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 [18] were added at three levels: 0, 6 and 12 dB. We used the HTK-toolkit [19] to train the HMM/GMM system with 80 context-dependent phonemes, 3 emitting states per phoneme and 12 mixtures per state.

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. Block-wise processing simply consists in (1) splitting a file in small blocks (e.g. 0.25 s), (2) each block is processed with an equal-weight strategy, as in [11]. EM fitting is done on 200 representative samples: 100 samples of the current block and 100 samples of the previous block.

Results are reported in Tab. 1, along with those of the baseline defined in Sect. 4.1. First of all, we note that the "ML" method systematically provides results inferior to those of the "moment" method. After looking at the parameters of the fitted models in both cases, we found that the "moment" method systematically yields larger σ_{I} values. σ_{I} is the threshold of the spectral subtraction. Hence, the "moment" method is more conservative: it leads to more noise removal, which may compensates some imperfections of the 2-mixture model. All other results and further discussions use the "moment" method only.

Results: Overall, the proposed approach provides a major improvement on both stationary (Lynx) and non-stationary (Factory) noise conditions, *without loss in clean conditions.*

As expected, block-wise processing leads to an additional improvement in non-stationary conditions (Factory), for the cost of a slight degradation in stationary conditions (Lynx and clean).

SNR (dB)	Subway	Babble	Car	Exhibition	Average
clean	**0.8**	**1.1**	**1.2**	**1.0**	**1.0**
20	3.5	3.0	3.6	4.1	3.5
15	6.8	6.8	7.3	7.4	7.1
10	14.7	14.1	16.0	15.2	15.0
5	32.2	37.1	34.5	30.9	33.7
0	64.5	74.1	69.0	59.5	66.8
-5	87.1	89.7	90.1	84.8	88.0
Average (0 to 20)	24.3	27.0	26.1	23.4	25.2

(a) Baseline

SNR (dB)	Subway	Babble	Car	Exhibition	Average
clean	1.2	1.4	1.3	1.1	1.3
20	**3.2**	**1.8**	**2.2**	**3.5**	**2.7**
15	**5.7**	**3.3**	**3.4**	**5.7**	**4.5**
10	**11.5**	**7.5**	**7.7**	**13.3**	**10.0**
5	**29.0**	**24.1**	**18.8**	**29.4**	**25.3**
0	**62.4**	**57.7**	**43.8**	**58.0**	**55.5**
-5	**85.3**	**86.4**	**78.8**	**83.0**	**83.4**
Average (0 to 20)	**22.3**	**18.9**	**15.2**	**22.0**	**19.6**

(b) Proposed approach (moment, 1-second blocks)

Table 2: WER results on Aurora 2, test set A. In all cases, both cepstral mean and variance normalization were applied. **Bold face** indicates the best result for each (condition, SNR) pair.

4.4 Experimental Results on Aurora 2 Task

All "try and test" experiments were done on the OGI Numbers 95 task, with final results reported and discussed in the previous section. In order to further confirm the obtained results, we picked one of the best approach (moment) and applied it "as is" on a larger, different task: Aurora 2 [16]). 1-second block-wise processing was chosen, since a few files in Aurora are much longer than the OGI Numbers 95 files.

The Aurora 2 task was designed to evaluate the front-end of ASR systems in noisy conditions [16]. The task is speaker-independent connected digit recognition. The database comprises isolated digits and sequences of up to 7 digits from the TIDigits database [20] spoken by male and female US-American adults. The original 20kHz data was downsampled to 8 kHz, in order to obtain a telephone bandpass between 0 and 4 kHz. The resulting data constitutes the clean speech data (clean condition). Noises were then added artificially at different SNR levels (20 dB to -5 dB). The noises were recorded at different places: suburban train, crowd of people (babble), car, exhibition hall, restaurant, street, airport, and train station. Unlike our OGI Numbers 95 experiments, nothing can be said strictly about the stationarity of the noise signals. Some noises are fairly stationary, for instance car noise and exhibition noise. Others contain non-stationary segments, e.g. street noise, babble noise and airport noise.

The Aurora 2 task defines two different training modes: training on clean condition only, and training on both clean and noisy conditions. In this paper, only experiments with training on clean condition are reported, because the purpose of our approach is precisely to remove noise as much as possible in order to make acoustic modeling as noise-robust as possible, *while retaining similar performance in clean conditions*. The details about the training set, test set and HTK recognizer can be found in [16].

Results on Aurora 2 test sets A and B with training on clean data are reported in Tab. 2 and 3,

SNR (dB)	Restaurant	Street	Airport	Train-stat.	Average
clean	**0.8**	**1.1**	**1.2**	**1.0**	**1.0**
20	2.6	3.3	2.8	3.8	3.1
15	5.2	6.3	5.5	7.9	6.2
10	12.2	14.2	12.3	16.3	13.7
5	31.2	32.7	28.8	35.2	32.0
0	65.2	63.4	61.8	69.9	65.1
-5	87.5	86.3	86.8	89.8	87.6
Average (0 to 20)	23.3	24.0	22.2	26.6	24.0

(a) Baseline

SNR (dB)	Restaurant	Street	Airport	Train-stat.	Average
clean	1.2	1.4	1.3	1.1	1.3
20	**1.6**	**2.4**	**1.8**	**1.7**	**1.9**
15	**4.2**	**4.6**	**2.9**	**4.2**	**4.0**
10	**9.2**	**9.8**	**6.1**	**9.3**	**8.6**
5	**26.5**	**24.0**	**17.8**	**20.8**	**22.3**
0	**58.9**	**50.7**	**44.5**	**47.7**	**50.4**
-5	**84.2**	**78.0**	**77.4**	**78.3**	**79.5**
Average (0 to 20)	**20.1**	**18.3**	**14.6**	**16.8**	**17.4**

(b) Proposed Approach (moment, 1-second blocks)

Table 3: WER results on Aurora 2, test set B. In all cases, both cepstral mean and variance normalization were applied. **Bold face** indicates the best result for each (condition, SNR) pair.

respectively. All Aurora 2 experiments used 10 ms frame shift and 25 ms frame length.

We did not run experiments on test set C, because it is the result of transmission through a channel with a “rising slope” frequency response, as described in [16]. This violates our assumption that the noise model is independent of f. A simple workaround would be to implement noise modeling and subtraction within each narrowband or critical band separately. The same approach could be used, but this is beyond the scope of this paper.

Finally, we need to mention that we also ran the baseline experiments without cepstral variance normalization, reported in Tab. 4, in order to check whether our baseline compares to the ones used in [16] and ICSLP 2002. Indeed, all results are similar, and average results are equal or better.

Results: From both Tab. 2 and 3, the conclusion is the same. For an average loss of 0.3% absolute in clean conditions, a major improvement is obtained in all noise conditions. The improvement is particularly large for 0, 5, 10 and 15 dB.

4.5 Discussion

To conclude, after initial tests on the OGI Numbers 95 task, we “froze” the approach, tested it on the Aurora 2 task. In both cases the exact same conclusion was drawn: a major improvement in noisy conditions can be obtained with the proposed pre-processing, while similar performances in clean conditions are retained. This is particularly interesting, given that (1) the modeling involves 4 parameters only, (2) fitting is fully unsupervised and does not require any tuning. Only the chosen duration of the block-wise processing can have an impact on non-stationary conditions.

5 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 models large magnitudes only. The 2-mixture model is trained on observed data in a fully unsupervised manner, using the EM algorithm. Obtained parameters are the basis for noise removal at the magnitude spectrogram level.

The "Unsupervised Spectral Subtraction" proposed in this paper is a simplification of our previous work [14]. The computational cost is very low, and in constrast to other works, no parameter tuning is involved. Experimental studies were conducted on both OGI Numbers 95 and Aurora 2 tasks, with the same conclusion: when applied as a pre-processing to MFCCs, the proposed approach allows for a major improvement in noise conditions, while retaining similar performances in clean conditions.

One direction for future work is large vocabulary conversational speech recognition. Another direction is application of the same type of modelling approach to a very different problem: microphone array-based speaker detection and localization. Promising results were reported in [14].

6 Acknowledgements

The authors acknowledge the support of the European Union through the AMI and HOARSE projects. This work was also carried out in the framework of the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2.

7 Annex A

In this section we derive the Rayleigh magnitude-domain silence model of $|\mathcal{F}_f^Y|$ (Eq. 1) from a white Gaussian assumption for the pre-emphasized signal Y_t.

First, let us recall a result originally shown by Rice in 1944 (for a demonstration see [12, 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:

$$f_R(r) = \frac{r}{\sigma} e^{-\frac{r^2}{2\sigma^2}} \quad \text{for} \quad r > 0 \tag{5}$$

Let us now define $Y_{1:N}$, as a vector of N uncorrelated[1] zero-mean Gaussian random variable $Y_{1:N} \overset{\text{def}}{=} [Y_1 \dots Y_N]^{\mathrm{T}}$. 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 $f = 1 \dots N$:

$$\mathcal{F}_f^Y \overset{\text{def}}{=} \sum_{n=1}^{N} Y_n e^{-2\pi(f-1)\frac{n-1}{N}} \tag{6}$$

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:

$$\begin{cases} A_f = \sum_{n=1}^{N} Y_n \cos\left(-2\pi(f-1)\frac{n-1}{N}\right) \\ B_f = \sum_{n=1}^{N} Y_n \sin\left(-2\pi(f-1)\frac{n-1}{N}\right) \end{cases} \tag{7}$$

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 [21, p. 99] it is also a zero-mean, single Gaussian random variable with variance:

$$\begin{cases} \sigma_{A_f}^2 = \sigma^2 \sum_{n=1}^{N} \cos^2\left(2\pi\left(f-1\right)\frac{n-1}{N}\right) \\ \sigma_{B_f}^2 = \sigma^2 \sum_{n=1}^{N} \sin^2\left(2\pi\left(f-1\right)\frac{n-1}{N}\right). \end{cases} \tag{8}$$

SNR (dB)	Subway	Babble	Car	Exhibition	Average
clean	0.9	0.9	1.1	0.9	1.0
20	3.9	2.6	2.8	4.1	3.3
15	9.4	6.4	8.0	9.4	8.3
10	29.2	22.3	33.0	28.6	28.3
5	63.4	56.1	68.8	64.6	63.2
0	80.1	78.7	81.5	82.3	80.7
-5	87.6	88.5	89.3	90.5	89.0
Average (0 to 20)	37.2	33.2	38.8	37.8	36.8

(a) Test Set A

SNR (dB)	Restaurant	Street	Airport	Train-stat.	Average
clean	0.9	0.9	1.1	0.9	1.0
20	2.1	3.4	2.2	2.1	2.5
15	5.0	8.4	4.8	6.0	6.1
10	17.3	26.6	15.4	22.2	20.4
5	47.9	60.0	48.1	57.5	53.4
0	74.8	79.6	73.0	77.4	76.2
-5	87.2	88.9	85.2	87.9	87.3
Average (0 to 20)	29.4	35.6	28.7	33.0	31.7

(b) Test Set B

Table 4: Aurora 2: MFCC Baseline results without cepstral variance normalization.

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

$$\begin{cases} \sigma^2_{A_f} = \frac{\sigma^2}{2}\left(N + \sum_{n=1}^{N} \cos\left(4\pi\,(f-1)\,\frac{n-1}{N}\right)\right) \\ \sigma^2_{B_f} = \frac{\sigma^2}{2}\left(N - \sum_{n=1}^{N} \cos\left(4\pi\,(f-1)\,\frac{n-1}{N}\right)\right). \end{cases} \tag{9}$$

Let us now write the complex domain sum:

$$\sum_{n=1}^{N} e^{j4\pi(f-1)\frac{n-1}{N}} = \sum_{n=0}^{N-1} \alpha^n = \frac{1-\alpha^N}{1-\alpha} = 0, \tag{10}$$

because $\alpha^N = 1$, where $\alpha = e^{j4\pi\frac{f-1}{N}}$. (Since $1 < f \leq N$, $\alpha \neq 1$ and all terms in Eq. 10 are defined.) From Eqs. 9 and 10 we conclude that:

$$\sigma_{A_f} = \sigma_{B_f} = \sigma\sqrt{\frac{N}{2}}. \tag{11}$$

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

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

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 *Proc. of ASRU*, 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 *Proc. 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] J. Cohen, "Application of an auditory model to speech recognition," *Journal of the Acoustic Society of America*, vol. 85, no. 6, June 1989.

[11] D. V. Campernolle, "Noise adaptation in a hidden markov model speech recognition system," *Computer Speech and Language*, vol. 13, no. 1, 1989.

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

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

[14] G. Lathoud, M. Magimai.-Doss, and B. Mesot, "A spectrogram model for enhanced source localization and noise-robust asr," in *Proceedings of Interspeech 2005*, 2005.

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

[16] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of ISCA ITRW ASR2000*, September 2000.

[17] . Lathoud, J. Odobez, and D. Gatica-Perez, "AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking," in *Proc. of the MLMI'04 Workshop*, 2005.

[18] 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.

[19] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 2.2*. Entropic Ltd, 1999.

[20] R. Leonard, "A database for speaker-independent digit recognition," in *Proceedings of ICASSP 2004*, 2004.

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