Spectral Intersections for Non-Stationary Signal Separation

Trausti Kristjansson[1], *Thad Hughes*[2]

[1]School of Science and Engineering, University of Reykjavik, Reykjavik, Iceland
[2]Google Inc., Mountain View CA, USA

traustithor@hr.is, thadh@google.com

Abstract

We describe a new method for non-stationary noise suppression that is simple to implement yet has performance that rivals far more complex algorithms.

Spectral Intersections is a model based MMSE signal separation method that uses a new approximation to the observation likelihood. Furthermore, Spectral Intersections uses an efficient approximation to the expectation integral of the MMSE estimate that could be described as *unscented importance sampling*.

We apply the new method to the task of separating speech mixed with music. We report results on the Google Voice Search task where the new method provides a 7% relative reduction in WER at 10 dB SNR. Interestingly, the new method provides considerably greater reduction in average WER than the Max method and approaches the performance of the more complex Algonquin algorithm.

Index Terms: Speech Recognition, Noise Robustness, Noise Suppression, Spectral Subtraction, Algonquin[1].

1. Introduction

With the rapid growth of speech recognition for mobile applications there is a greater need for robustness to non-stationary noise interference such as music. Traditional methods such as Spectral Subtraction[1] and Ephraim Malah[2] are effective for suppression of stationary noise, but are ill suited for non-stationary noise. Model based methods such as Max and Algonquin can perform well [3] for very non-stationary noise such as music, but are complex. A good overview of the state of the art of model based methods is provided by Hershey et al. [4].

Spectral Intersections is in the family of model based Minimum Mean Squared Error (MMSE) algorithms such as Wiener, Max[5] and Algonquin[6][7]. It offers good performance while being easy to implement and is based on a new and interesting approximation.

1.1. Background and Overview

In this work, we model the signals in the log spectrum domain. Starting with the time domain signal $\mathrm{x}[t]$, we first compute the short time Fourier transform of sequential segments of the signal $X(f)$ and finally the log spectrum $x = \log(|X(f)|)$.

We use Gaussian Mixture Models (GMMs) for the component signals in the log spectrum domain

$$p(x) = \sum_i p_i(x) = \sum_i \pi_i N(x; \mu_i, \Sigma_i). \tag{1}$$

where π_i is the mixture weight, μ_i is the mixture mean, and Σ_i is the covariance matrix for mixture i.

The observed signal y_{obs} is an acoustic mixture of the target signal x_1 and interference signal x_2. The MMSE estimate for the separated target signal x_1 is:

$$\begin{aligned}\hat{x}_1 = \mathrm{E}[x_1|y_{obs}] &= z \int x_1 \cdot p(y_{obs}, x_1, x_2) dx_1 dx_2, \\ &= z \sum_{i,j} \int x_1 \cdot p(y_{obs}|x_1, x_2) p_i(x_1) p_j(x_2) dx_1 dx_2,\end{aligned} \tag{2}$$

where $p(y_{obs}|x_1, x_2)$ is the observation likelihood, $p(x_1)$ is the target prior model, $p(x_2)$ is the interference prior model and $z = p(y_{obs})^{-1}$. In this paper we will interchangeably talk about the speech model for the target model and music model for the interference model.

In Section 2 we explain the method for approximating the observation likelihood by taking only the constructive and destructive combinations into account. In Section 3 we explain a lightweight approximation to the expectation integral required for the MMSE estimate. In Section 4 we report results on the real world task of recognizing Voice Search utterances mixed with music and in Section 5 we summarize and contrast the new method to Algonquin and Max methods.

2. Signal Mixing Model

The model for mixed speech and music in the time domain is

$$\mathrm{y}[t] = \mathrm{x}_1[t] + \mathrm{x}_2[t] \tag{3}$$

[1] This work was performed while Trausti Kristjansson was at Google Research.

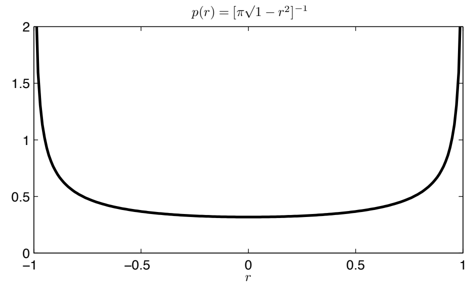

Figure 1: The plot shows the distribution for $r = \cos(\theta)$, i.e. $p(r) = [\pi\sqrt{1-r^2}]^{-1}$. Notice the peaks at either end of the interval $(-1, 1)$.

where $\mathrm{x}_1[t]$ denotes the target speech signal at time t, $\mathrm{x}_2[t]$ denotes the interference music signal and $\mathrm{y}[t]$ denotes the mixed signal. In the Fourier domain, the relationship becomes

$$Y(f) = X_1(f) + X_2(f) \tag{4}$$

where f designates the frequency component of the FFT. This can also be written in terms of the magnitude and the phase of each component:

$$|Y(f)|\angle Y(f) = |X_1(f)|\angle X_1(f) + |X_2(f)|\angle X_2(f) \tag{5}$$

where $|Y(f)|$ is the magnitude of $Y(f)$ and $\angle Y(f)$ is the phase and similarly for X_1 and X_2.

2.1. Constructive and Destructive Mixing

By the law of cosines, the relationship between the components is (dropping the dependence on the frequency f)

$$|Y|^2 = |X_1|^2 + |X_2|^2 + 2|X_1||X_2|\cos(\theta) \tag{6}$$

where θ is the angle between X_1 and X_2. The target and interference signals are independent and it follows that θ can be treated as a uniformly distributed random variable. The distribution for $r = \cos(\theta)$ is $p(r) = [\pi\sqrt{1-r^2}]^{-1}$ which has sharp peaks at either end of the interval $(-1, 1)$ as shown in Figure 1. These correspond to cases where the component signals are exactly in phase or exactly out of phase or $\theta = \{0, \pi\}$, i.e. the *constructive* and *destructive* combinations respectively. This suggests that we can reasonably approximate this distribution by considering only the constructive and destructive combinations.

If the acoustic mixing is exactly constructive, Equation (5) becomes

$$|Y| = |X_1| + |X_2| \tag{7}$$

and if the acoustic mixing is exactly destructive, we have two cases

$$|Y| = |X_1| - |X_2| \quad \text{if } |X_1| > |X_2| \tag{8}$$

$$|Y| = |X_2| - |X_1| \quad \text{if } |X_2| > |X_1|. \tag{9}$$

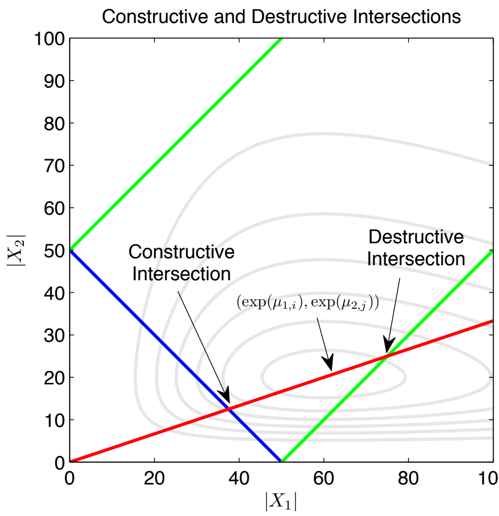

Figure 2: Given Y, the plot shows the constructive (blue) $|Y| = |X_1| + |X_2|$ and destructive (green) $|Y| = |X_1| - |X_2|$ and $|Y| = |X_2| - |X_1|$ lines. Also shown are the contours of the prior distribution for a particular mixture combination i, j with mode $[\mu_{1,i}, \mu_{2,j}]$ in the log spectrum domain corresponding to $(\exp(\mu_{1,i}), \exp(\mu_{2,i}))$ in the magnitude spectrum domain.

These equations describe valid combinations of $|X_1|$ and $|X_2|$. In Figure 2 the constructive combinations are shown as the blue downward line and the destructive combinations are shown as the green lines.

2.2. Constructive and Destructive Intersections

Recall that we use Gaussian mixture models for the prior models for the target signal $p(x_1)$ and interference signals $p(x_2)$. Consider a particular mixture combination i, j where i is the index for the target mixture and j is the index of interference mixture. The joint distribution for this combination, $p_{i,j}(x_1, x_2)$, is a product of two Gaussian distributions with a mode $[\mu_{1,i}, \mu_{2,j}]$ in the log spectrum domain. In the magnitude spectrum domain, we denote this coordinate as $(\exp(\mu_{1,i}), \exp(\mu_{2,j}))$

Consider a line from the origin $(0, 0)$ through the coordinate $(\exp(\mu_{1,i}), \exp(\mu_{2,i}))$. This line is shown in Figure 2 as the red line.

Now consider the intersection of this line to the lines for the constructive and destructive mixing. We call the intersection points the *constructive intersection* and the *destructive intersection*. In the magnitude spectrum domain, the coordinate of the constructive intersection point is $(|X^{con}_{1,i,j}|, |X^{con}_{2,i,j}|)$ where

$$|X^{con}_{1,i,j}| = \frac{|Y|}{1 + \exp(\mu_{2,j})/\exp(\mu_{1,i})} \tag{10}$$

$$|X^{con}_{2,i,j}| = \frac{|Y|}{1 + \exp(\mu_{1,i})/\exp(\mu_{2,j})} \tag{11}$$

and the coordinate of the destructive intersection point is $(|X^{des}_{1,i,j}|, |X^{des}_{2,i,j}|)$ where

$$|X^{des}_{1,i,j}| = \frac{|Y|}{1 - \exp(\mu_{2,j})/\exp(\mu_{1,i})} \tag{12}$$

$$|X^{des}_{2,i,j}| = \frac{|Y|}{-1 + \exp(\mu_{1,i})/\exp(\mu_{2,j})}, \tag{13}$$

for the case $|X_1| > |X_2|$ and similarly with signs reversed for the case $|X_2| > |X_1|$. In the log spectrum domain the constructive intersection is

$$(x^{con}_{1,i,j}, x^{con}_{2,i,j}) = (\log(|X^{con}_{1,i,j}|), \log(|X^{con}_{2,i,j}|)) \tag{14}$$

and similarly for the destructive intersection.

3. Unscented MMSE

Since we use Gaussian mixture models for the prior models, the exact evaluation of the MMSE estimate requires that we evaluate an integral for each combination i, j in Equation (2).

Instead of computing the full integral over x_1 and x_2, we use an approximation that can be seen as an extremely sparse *importance sampling*[8] method for finding the expectation in Equation (2). Instead of using random samples drawn from a proposal distribution, we use only two samples, i.e. the samples corresponding to the constructive and destructive intersections.

The *spectral intersection* estimate for the target signal is

$$\begin{aligned}\hat{x}_1 = \mathrm{E}[x_1|y_{obs}] \approx Z \cdot \sum_{i,j} \pi_i \pi_j \cdot \\ \cdot \{x^{con}_{1,i,j} \cdot N(x^{con}_{1,i,j}, \mu_{1,i}, \Sigma_{1,i}) N(x^{con}_{2,i,j}, \mu_{2,j}, \Sigma_{2,j}) \\ + x^{des}_{1,i,j} \cdot N(x^{des}_{1,i,j}, \mu_{1,i}, \Sigma_{1,i}) N(x^{des}_{2,i,j}, \mu_{2,j}, \Sigma_{2,j})\}\end{aligned} \tag{15}$$

where Z is a normalizing factor[2]. Notice that this amounts to little more than evaluating the prior distribution at the constructive and destructive intersections.

The way in which we chose the sample points is reminiscent of the *unscented approximation* [9]. Hence, this method could be called *unscented importance sampling*.

4. ASR Experiments

We evaluated the algorithms on real utterances from the Google Voice Search system. Since the utterances are generally near field and of high SNR, we artificially added music to the data to produce noise conditions with varying SNR. We evaluated using our state of the art ASR system. A more detailed explanation of the system and discussion of the influence of the size of the music model for Max and Algonquin can be found in [3].

4.1. Dataset characteristics

The dataset consists of approximately 38,000 manually-transcribed utterances containing 38 hours of anonymized English-language spoken queries to Google Voice Search. The utterances were spoken by 296 different speakers, and range in length from 0.2 to 12.3 seconds, with a mean of 3.6 seconds. The utterances were recorded and stored in 16-bit, 16kHz uncompressed format.

The dataset contains a varying amount of speech for each speaker and hence the amount of training data for each speech model is different.

4.2. Training speech models

To train the speech model for each speaker the data was segmented into a low-noise training data and higher noise-test data. We compute 256-dimensional log-spectral feature vectors for each of the speaker's utterances, using 25ms frames spaced at 10ms intervals.

Since this is real data, much of the *cleaner* speech data contains low non-stationary background noise, such as TV noise. If speech models are trained directly on this data, the majority of the model components are allocated to modeling this low non-stationary noise background. To circumvent this problem, we use a percentile based VAD to separate the low-noise condition speech into speech frames and non-speech frames.

From the speech frames, we estimate a GMM with at most 200 components subject to the constraint that there are at least 20 frames per Gaussian component. From the non-speech frames we estimate a smaller 20-component GMM. These two models are then combined to form a clean-speech GMM.

4.3. Training noise models

At least 30% of the data for each speaker is held out as test data. For each utterance in the test set, we select a random song from a database of 500 popular songs, and mix it with the utterance at the desired SNR. Noise models are trained on the music directly prior to the speech. Hence we included 8 seconds of musical prologue before the onset of speech in the utterance. We then compute the same 256-dimensional log-spectral feature vectors used to create the speech model, and use the feature frames from the prologue to construct 8 mixture noise GMMs.

4.4. Experimental Setup

The SNR of the utterances for each speaker is first computed. Based on the SNR, they are divided into training and testing sets, where the least-noisy 70% of the data is used for training and the remaining 30% is used for testing.

[2] $Z^{-1} = \sum_{i,j,k} \pi_i \pi_j N(x^k_{1,i,j}, \mu_{1,i}, \Sigma_{1,i}) N(x^k_{2,i,j}, \mu_{2,j}, \Sigma_{2,j})$

	Max	SI	Algon.
Average WER reduction	2.7	4.0	4.8
WER reduction at 10 dB	4.6	7.6	7.0

Table 1: Average reduction in WER for all conditions, 10 dB - 20 dB.

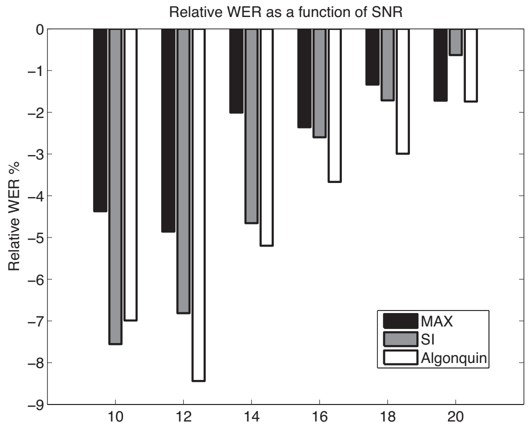

Figure 3: Comparison of WER results for the Max, Spectral Intersection and Algonquin Algorithms for a range of SNRs.

4.5. Signal separation and evaluation

We then apply the Max, Algonquin and Spectral Intersections noise reduction techniques using the per-speaker speech model constructed from the speaker's training data, and the per-utterance noise model constructed from each utterance's prologue.

The resulting cleaned feature frame sequence is then re-synthesized as a waveform using the overlap-add algorithm and sent to the speech recognizer to test the denoising quality. All speech recognition was performed with a recent version of Google's Voice Search speech recognizer. This system uses an acoustic model with approximately 8000 context-dependent states and approximately 330k Gaussians, and is trained with LDA, STC, and MMI. The Voice Search language model used for recognition contains more than one million English words.

We did not retrain the acoustic models on denoised data which would be expected to give better results. However, the relative performance of the respective methods is expected to remain the same.

4.6. Results

The average performance across noise conditions is a 4% relative reduction in error rate which approaches the performance of Algonquin while substantially outperforming the Max method.

As can be seen from Table 1 and Figure 3 the Spectral Intersections method follows the trend of Algonquin but provides slightly less gain in all conditions except the noisiest 10 dB condition where it outperforms Algonquin. However, it exceeds Max in almost all conditions.

5. Discussion

We have presented a new method for non-stationary noise suppression. The method is easy to implement but has performance that rivals far more complex methods.

The computational complexity of the new method is the same as that of Max or Algonquin, which is dominated by the cross product of the number of mixtures in the target and interference models, i.e. $O(I \cdot J)$, where I is the number of mixtures in the target model and J is the number of mixtures in the interference model. However, Algonquin requires Newton iterations of a Laplace transform involving a matrix inverse and Max requires the computation of the log of the cumulative normal distribution which requires careful attention to numerical stability when implementing.

In contrast, the new method involves simple line intersections and the evaluation of standard GMM distributions which are well understood and optimized.

6. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," in *IEEE Trans. on Acoust., Speech, Signal Processing, vol. ASSP-33*, 1985, pp. 443–445.

[3] T. Hughes and T. Kristjansson, "Music models for music-speech separation," in *Proceedings of ICASSP*, 2012.

[4] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition : A graphical modeling approach," in *Computer Speech and Language*, 2009.

[5] A. Nadas, Nahamoo, and M. D., Picheny, "Speech recognition using noise-adaptive prototypes," in *Proceedings of ICASSP*, 1989, pp. 1495 –?1503.

[6] B. Frey, L. Deng, A. Acero, and T. Kristjansson, "Algonquin: Iterating laplace's method to remove multiple types of acoustic distortion for robust speech recognition," in *Proceedings of Eurospeech*, 2001.

[7] J. Droppo, L. Deng, and A. Acero, "A comparison of three non-linear observation models for noisy speech features," in *Proceedings of Eurospeech*, 2003.

[8] S. B. M. Ferrari, "Importance sampling simulation of turbo product codes," in *ICC2001, The IEEE International Conference on Communications*, 2001, pp. 2773–2777.

[9] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte, "A new approach for filtering nonlinear systems," in *Proceedings of the American Control Conference*, 1995, pp. 1628 – 1632.