A NEW APPROACH TO SPEECH ENHANCEMENT BY A MICROPHONE ARRAY USING EM AND MIXTURE MODELS

Hagai Attias, Li Deng

{hagaia,deng}@microsoft.com
Microsoft Research
1 Microsoft Way, Redmond, WA 98052, USA

ABSTRACT

Speech enhancement and recognition in noisy, reverberant conditions is a challenging open problem. We present a new approach to this problem, which is developed in the framework of probabilistic modeling. Our approach incorporates information about the statistical structure of speech signals using a speech model, which is pre-trained on a large dataset of clean speech. The speech model is a component in a larger model describing the observed sensor signals. That model is parametrized by the coefficients of the reverberation filters and the spectra of the sensor noise. We develop an EM algorithm that estimates those parameters from data and constructs a Bayes optimal estimator of the original speech signal.

1. INTRODUCTION

Speech enhancement in a realistic environment is a challenging problem, which remains unsolved after more than three decades of research. A successful technique would find many applications, particularly in the domains of speech recognition, natural user interfaces, and communications. The difficulty of the enhancement task depends strongly on environmental conditions. When a speaker is close to a microphone and the noise level is low, reverberation effects are fairly small and standard signal processing techniques yield satisfactory performance. However, as the distance from the microphone increases, the distortion of the speech signal, resulting from large amounts of noise and significant reverberation, becomes gradually more severe.

Much work has been done on speech enhancement using traditional signal processing methods, such as spectral subtraction, noise cancellation, and array processing (see, e.g., [5]). Whereas these methods have had many well known successes, they have so far fallen short of offering a satisfactory, robust solution to the enhancement problem. One shortcoming of these methods is that they typically exploit just second order statistics, i.e., correlation functions or spectra, of the sensor signals, ignoring higher order statistics. In other words, they implicitly make a Gaussian assumption on speech signals that are highly non-Gaussian. A related issue is that these methods typically disregard information on the statistical structure of speech signals. In addition, some of these methods suffer from the lack of a principled framework. This is sometime manifested in ad-hoc solutions, e.g., spectral subtraction algorithms recover the speech spectrum of a given frame by essentially subtracting the noise spectrum from the sensor signal spectrum, requiring a special treatment when the result is negative. Another example is the difficulty of combining algorithms that remove noise with algorithms that handle reverberation into a single system in a systematic manner.

More recently, a new type of speech enhancement algorithms have started to emerge [3,1,4]. These algorithms follow from taking a probabilistic modeling approach to the problem. In that approach, one starts by constructing a model for clean speech signals. This model, usually a mixture model or a hidden Markov model (HMM), is trained offline on a large dataset of clean speech. The speech model is combined with a similar (usually much smaller) noise model to create a model for the observed sensor signals. The resulting model is a *hidden variable model*, where the original speech signal and speech state are the hidden (unobserved) variables, and the sensor signals are the data (observed) variables. As usual in hidden variable models, a suitable EM algorithm estimates from data the noise parameters and the original speech signal.

The probabilistic modeling approach has several advantages. Parameter and signal estimators benefit from the optimality properties of maximum likelihood. These estimators are derived in a principled fashion and no ad-hoc procedures are needed. Information about the structure of speech signals can be incorporated using a speech model. And due to the mixture form of that model, higher order statistics of the sensor signals are taken into account automatically when reconstructing the original speech signal. This approach has been applied successfully to noise removal using a single microphone which is located near the speaker [3,1,4].

However, these methods have not yet been extended to

the case where the speaker is far away from the microphone. One reason this extension is difficult is that in such cases, distortions originate not just from noise but also from reverberation effects. Reverberations typically operate on a longer time scale compared to the time scale of the speech model, creating long range temporal correlations in the data. It has been unclear how these correlations could be described in the probabilistic modeling framework. Furthermore, these correlations were expected to cause intractability of the resulting models, which would be likely to hamper such a treatment.

This paper significantly extends the scope of the probabilistic modeling approach to speech enhancement. We develop this approach for the case where the speech signal is distorted by both noise and reverberation. To overcome intractabilities arising from reverberation-induced temporal correlations, we develop an efficient approximation scheme that reduces the computational complexity to essentially that of the noise-only case. In addition, we extend the probabilistic modeling approach beyond a single sensor, deriving an enhancement algorithm that can take advantage of a microphone array. This is a EM algorithm, where the M-step updates the parameters of the noise signals and reverberation filters, and the E-step updates the sufficient statistics, which include the speech signal estimator.

2. NOISY REVERBERANT SPEECH AND SUBBAND FILTERING

We start with a mathematical description of the problem. Let $x[n]$ denote the source signal at time point n, and let $y^i[n]$ denote the signal received at sensor i at the same time. As it propagates toward the sensors, the source signal is distorted by several factors, including the response of the propagation medium and multipath propagation conditions. The resulting reverberation effects may be modeled quite accurately by linear filters applied to the source signal. Background noise and sensor noise, which are assumed to be additive, lead to additional distortion. Hence we have

$$y^i[n] = \sum_m h^i[m]x[n-m] + u^i[n] \ , \tag{1}$$

where $h^i[m]$ denotes the impulse response of the filter corresponding to sensor i, and $u^i[n]$ is the associated noise.

Rather than time domain signals like $x[n]$ above, we will be working throughout the paper with subband signals. These signals are obtained by applying an N-point window to the signal at equally spaced points, and computing the FFT of the windowed signal. For the speech signal $x[n]$, let $X_m[k]$ denote the mth subband signal, also termed *frame*, defined by

$$X_m[k] = \sum_n e^{-i\omega_k n} w[n]x[mJ+n] \tag{2}$$

where $w[n]$ is the window function, which vanishes outside $n \in [0, N-1]$, and $J > 0$ is the spacing between the starting points of the windows. $k = (0 : N-1)$ runs over the subbands, and $m = (0 : M-1)$ indexes the frames. The subband signals $Y^i_m[k]$ and $U^i_m[k]$ corresponding to the sensor and noise signals can also be defined.

We will assume that the subband signals satisfy the same relation as the time domain signals (1),

$$Y^i_m[k] = \sum_n H^i_n[k]X_{m-n}[k] + U^i_m[k] \tag{3}$$

where the complex quantities $H^i_n[k]$ are related to the filters $h^i[m]$ by a linear transformation whose exact form is omitted. Of course, the relation (3) is exact only in the limit $N \to \infty$. For finite N it is approximate, but but the approximation turns out to be quite accurate for a suitable choice of the window function.

3. PROBABILISTIC SIGNAL MODELS

We now define the models used in this paper.

Notation. For a complex variable Z, we define a Gaussian distribution with mean μ and precision (defined as the inverse variance) ν by

$$p(Z) = \mathcal{N}(Z \mid \mu, \nu) = \frac{\nu}{\pi} \exp(-\nu \mid Z - \mu \mid^2) \ . \tag{4}$$

Viewed as a joint distribution over $\mathrm{Re}Z$ and $\mathrm{Im}Z$, $p(Z)$ integrates to one, and satisfies $E(Z) = \mu$, $E(\mid Z \mid^2) = \mid \mu \mid^2 + 1/\nu$. The operator E denotes averaging.

When building statistical models of subband signals, we will ignore the real-valued subbands $k = 0, N/2$ and focus on the complex ones. We define the complex $(N/2-1)$-dim vector X_m containing all subbands of frame m,

$$X_m = (X_m[1], ..., X_m[N/2-1]) \tag{5}$$

(for $k > N/2$, $X_m[k] = X_m[N-k]^\star$). We also let $X[k]$ denote subband k of all frames, and let X denote all subbands of all frames,

$$\begin{aligned} X[k] &= \{X_m[k], m = (0 : M-1)\} \ , \\ X &= \{X_m[k], k = (0 : N-1), m = (0 : M-1)\} \ . \end{aligned} \tag{6}$$

A corresponding notation is used for Y^i and U^i.

Mixture model for speech. We describe the speech frames X_m by a C-component Gaussian mixture model. S_m denotes the component label at frame m, which assumes the value $s = (1 : C)$ with probability π_s. Component s has mean zero and precision A_s. Hence

$$\begin{aligned} p(X_m \mid S_m = s) &= \prod_{k=1}^{N/2-1} \mathcal{N}(X_m[k] \mid 0, A_s[k]) \ , \\ p(S_m = s) &= \pi_s \ . \end{aligned} \tag{7}$$

This Gaussian has a diagonal covariance matrix with $1/A_s[k]$ on the diagonal, leading to the interpretation of the precisions as the inverse spectrum of component s, since

$$E(\mid X_m[k] \mid^2 \mid S_m = s) = 1/A_s[k] \ . \tag{8}$$

For X_m we thus have the mixture distribution $p(X_m) = \sum_s p(X_m \mid S_m = s)p(S_m = s)$. Notice that, whereas different subbands of a given component are independent, subbands of X_m are correlated via the summation over components.

Assuming i.i.d. frames, we have

$$p(X \mid S) = \prod_m p(X_m \mid S_m) \ , \quad p(S) = \prod_m p(S_m) \tag{9}$$

where S denotes the labels in all frames collectively, $S = \{S_m, m = (0 : M-1)\}$. This specifies our speech model, parametrized by $\{A_s, \pi_s\}$. In actual speech signals the frames are not i.i.d.; speech recognition engines commonly use HMMs to describe inter-frame correlations. It is straightforward to incorporate such models into our framework, but, as this paper is introducing the framework, we prefer to stick with the simplifying i.i.d. assumption.

Noise model. For the noise recorded at sensor i, we use a colored zero-mean Gaussian model with spectrum $1/B^i[k]$,

$$p(U^i_m) = \prod_k \mathcal{N}(U^i_m[k] \mid 0, B^i[k]) \ . \tag{10}$$

We assume the that noise signals at difference sensors are uncorrelated, but this assumption can be easily relaxed. (Notice that noise cancellation algorithms actually rely on noise correlation between sensors.) From the i.i.d. assumption we get $p(U^i) = \prod_m p(U^i_m)$.

Conditional sensor distribution. The noise model implies the distribution of the sensor signals conditioned on the original speech signal. Using (3), we substitute $U^i_m[k] = Y^i_m[k] - \sum_n H^i_n[k]X_{m-n}[k]$ in (10) and obtain

$$p(Y^i_m \mid X) = \prod_k \mathcal{N}(Y^i_m[k] \mid \sum_n H^i_n[k]X_{m-n}[k], B^i[k]) \ , \tag{11}$$

where $X = \{X_m[k]\}$ as defined above. Note that the sensor signal distribution at frame m depends on the speech signal at the same frame but also at previous frames. The noise frames being i.i.d. lead to

$$p(Y^i \mid X) = \prod_m p(Y^i_m \mid X) \ . \tag{12}$$

Complete data distribution. The complete data comprise the observed variables $Y = \{Y^i\}$ and the unobserved ones X, S. Using the assumption of sensor independence, we get the complete data distribution in our model,

$$p(Y, X, S) = \prod_i p(Y^i \mid X)p(X \mid S)p(S) \ , \tag{13}$$

whose factors are specified by (9) and (12).

Training. The speech model is trained offline on a large speech database including 150 male and female speakers reading sentences from the Wall Street Journal (see [1] for details). The noise model is estimated either offline or from quiet moments in the noisy signal.

4. AN EM ALGORITHM

Here we develop an EM algorithm that estimates the filter parameters $H^i_m[k]$ and the noise spectra $B^i[k]$ from the data Y. It also computes the required sufficient statistics (SS) and the speech signal estimator $\hat{X}_m[k]$.

Unfortunately, a straightforward implementation of EM in our model leads to a computationally intractable algorithm. To see this, recall that the central object of the E-step is the conditional distribution over the unobserved variables X, S given the observed ones Y, $p(X, S \mid Y)$. This distribution, termed the *posterior distribution*, can in principle be obtained from the complete data distribution (13) via Bayes' rule. It is from the posterior that one derives the SS. The difficulty comes from having to sum over the C^M configurations of component labels $S = (S_0, ..., S_{M-1})$, where C is the number of speech model components and M the number of frames. Speech models that lead to good performance include at least 100 components. Whereas for short filters (i.e., relative to the window length N) $M = 1, 2$ and exact summation is possible, realistic scenarios have $M \geq 5$, which require summation over at least 10^{10} configurations.

In this paper, we present an EM algorithm that uses a systematic approximation to compute the SS. The effect of the approximation is to introduce an additional iterative procedure nested within the E-step. Our approximation is based on recent techniques developed in the field of machine learning, termed *variational techniques* [6]. Derivation details, as well as a proof of convergence, will be provided in a longer version of this paper.

Sufficient statistics. For each frame m and subband k, the E-step computes (1) the conditional mean and precision of $X_m[k]$ given $S_m = s$ and the observed data Y, denoted by $\rho_{sm}[k]$ and $\nu_{sm}[k]$, and (2) the conditional probability that $S_m = s$ given Y, denoted γ_{sm}. Mathematically,

$$\begin{aligned} \rho_{sm}[k] &= E(X_m[k] \mid S_m = s, Y) \ , \\ \nu_{sm}[k] &= E(\mid X_m[k] \mid^2 \mid S_m = s, Y) - \mid \rho_{sm}[k] \mid^2 \ , \\ \gamma_{sm} &= p(S_m = s \mid Y) \end{aligned} \tag{14}$$

where E denotes averaging w.r.t. $p(X_m[k] \mid S_m = s, Y)$.

These quantities are computed in the E-step below. Using them, we compute the mean of the speech signal $\hat{X}_m[k]$ conditioned on the data,

$$\hat{X}_m[k] = E(X_m[k] \mid Y) = \sum_s \gamma_{sm}\rho_{sm}[k] \ , \tag{15}$$

which serves as our speech estimator. We also compute its autocorrelation $\lambda_m[k]$, and its cross correlation with the data $\eta_m[k]$,

$$\begin{aligned}
\lambda_m[k] &= \sum_n E(X_{n+m}[k]X_n[k]^\star \mid Y) \ , \\
\lambda_{m>0}[k] &= \sum_n \hat{X}_{n+m}[k]\hat{X}_n[k]^\star \ , \\
\lambda_{m=0}[k] &= \sum_{sn} \gamma_{sn}(\mid \rho_{sn}[k] \mid^2 + \frac{1}{\nu_{sn}}) \ , \\
\eta^i_m[k] &= \sum_n E(Y^i_{n+m}[k]X_n[k]^\star \mid Y) \\
&= \sum_n Y^i_{n+m}[k]\hat{X}_n[k]^\star \ . \qquad (16)
\end{aligned}$$

These SS are used in the M-step.

M-step. Here we solve

$$\sum_n H^i_n[k]\lambda_{m-n}[k] = \eta^i_m[k] \qquad (17)$$

for $H^i_n[k]$. This can be done easily using subband FFT, as follows. For each subband k, define the M-point FFT of $H^i_m[k]$ by

$$\tilde{H}^i[k,l] = \sum_{m=0}^{M-1} e^{-i\tilde{\omega}_l m} H^i_m[k] \ , \qquad (18)$$

where $\tilde{\omega}_l = 2\pi l/M$ are the frequencies, $l = (0 : M-1)$. The subband FFTs $\tilde{\lambda}[k,l]$ and $\tilde{\eta}^i[k,l]$ are defined in the same manner. We then get

$$\tilde{H}^i[k,l] = \frac{\tilde{\eta}^i[k,l]}{\tilde{\lambda}[k,l]} \ . \qquad (19)$$

The update rule for the noise spectra are omitted.

E-step. The means $\rho_{sm}[k]$ (14) are obtained by solving

$$\begin{aligned}
\sum_{in} B^i[k]H^i_{n-m}[k]^\star(Y^i_n[k] - \sum_{r\neq m} H^i_{n-r}[k]\hat{X}_r) \\
= \nu_{sm}[k]\rho_{sm}[k] \ , \qquad (20)
\end{aligned}$$

where the variances are given by

$$\nu_{sm}[k] = \sum_{in} B^i[k] \mid H^i_{n-m}[k] \mid^2 + A_s[k] \ . \qquad (21)$$

Finally, the update rule for the probabilities γ_{sm} (14) is expressed in term of its logarithm,

$$\begin{aligned}
\log\gamma_{sm} &= \sum_k \left(\nu_{sm}[k] \mid \rho_{sm}[k] \mid^2 + \log\frac{A_s[k]}{\nu_{sm}[k]} \right) \\
&+ \log\pi_s \ . \qquad (22)
\end{aligned}$$

The E-step equations are solved iteratively, since the $\rho_{sm}[k]$ and the γ_{sm} are nonlinearly coupled.

5. EXPERIMENTS

In preliminary experiments, we have tested the algorithm using 5 sentences from the WSJ dataset, working at a 16kHz sampling rate. We used real room, 2000-tap filters, whose impulse responses have been measured separately using a microphone array. We also used noise signals recorded in an office containing a PC and A/C. For each sentence, two microphone signals were created by convolving it with two different filters and adding two noise signals at 10dB SNR. Using a pre-trained noise model, the algorithm estimated the filter parameters and the original speech signals from the microphone data. Averaged over sentences, the mean SNR in the estimated signal was 13.9dB.

6. CONCLUSIONS AND EXTENSIONS

We have presented and demonstrated a new framework for speech enhancement in noisy, reverberant situations, based on probabilistic modeling. While the experiments reported here focus on improvement in SNR, experiments now in progress test whether this translates into improvement in word recognition rates, by feeding the enhanced signal to a recognition engine (see [1] for the noise-only case). We are currently working on several extensions of this approach, including (1) An online version of our algorithm that tracks the parameters and adapt the speech estimator in non-stationary situations; (2) Handling multi-speaker cases (see, e.g., [2]) by incorporating multiple source models, which results in a novel source separation algorithm; (3) Wokring in the cepstral domain rather than in the frequency domain to form a direct connection with recogntion engines.

7. REFERENCES

[1] H. Attias, L. Deng, A. Acero, J.C. Platt (2001). A new method for speech denoising using probabilistic models for clean speech and for noise. *Proc. Eurospeech 2001*.

[2] H. Attias, C.E. Schreiner (1998). Blind source separation and deconvolution: the dynamic component analysis algorithm. *Neural Computation* 10, 1373-1424.

[3] Ephraim, Y. (1992). Statistical model based speech enhancement systems. *Proc. IEEE* 80(10), 1526-1555.

[4] B.J. Frey, L. Deng, A. Acero, T. Kristjansson (2001). An iterative variational method for removing multiple types of acoustic distortion for robust speech recognition. *Proc. Eurospeech 2001*.

[5] S. Griebel, M. Brandstein (2001). Microphone array speech dereverberation using coarse channel modeling. *Proc. ICASSP 2001*.

[6] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, L.K. Saul (1999). An introduction to variational methods in graphical models. *Machine Learning* 37, 183-233.