HIGH RESOLUTION SIGNAL RECONSTRUCTION

Trausti Kristjansson

Machine Learning and Applied Statistics
Microsoft Research
traustik@microsoft.com

John Hershey

University of California, San Diego
Machine Perception Lab
jhershey@cogsci.ucsd.edu

ABSTRACT

We present a framework for speech enhancement and robust speech recognition that exploits the harmonic structure of speech. We achieve substantial gains in signal to noise ratio (SNR) of enhanced speech as well as considerable gains in accuracy of automatic speech recognition in very noisy conditions.

The method exploits the harmonic structure of speech by employing a high frequency resolution speech model in the log-spectrum domain and reconstructs the signal from the estimated posteriors of the clean signal and the phases from the original noisy signal.

We achieve a gain in signal to noise ratio of 8.38 dB for enhancement of speech at 0 dB. We also present recognition results on the Aurora 2 data-set. At 0 dB SNR, we achieve a reduction of relative word error rate of 43.75% over the baseline, and 15.90% over the equivalent low-resolution algorithm.

1. INTRODUCTION

A long standing goal in speech enhancement and robust speech recognition has been to exploit the harmonic structure of speech to improve intelligibility and increase recognition accuracy.

The source-filter model of speech assumes that speech is produced by an excitation source (the vocal cords) which has strong regular harmonic structure during voiced phonemes. The overall shape of the spectrum is then formed by a filter (the vocal tract). In non-tonal languages the filter shape alone determines which phone component of a word is produced (see Figure 2). The source on the other hand introduces fine structure in the frequency spectrum that in many cases varies strongly among different utterances of the same phone.

This fact has traditionally inspired the use of smooth representations of the speech spectrum, such as the Mel-frequency cepstral coefficients, in an attempt to accurately estimate the filter component of speech in a way that is invariant to the non-phonetic effects of the excitation[1].

There are two observations that motivate the consideration of high frequency resolution modelling of speech for noise robust speech recognition and enhancement. First is the observation that most noise sources do not have harmonic structure similar to that of voiced speech. Hence, voiced speech sounds should be more easily distinguishable from environmental noise in a high dimensional signal space[1].

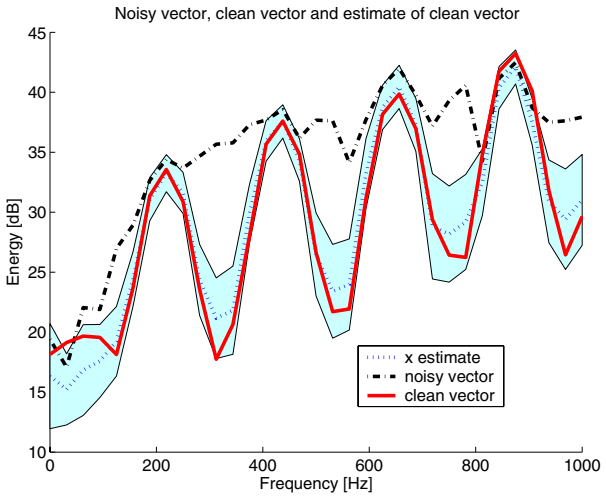


Fig. 1. The noisy input vector (dot-dash line), the corresponding clean vector (solid line) and the estimate of the clean speech (dotted line), with shaded area indicating the uncertainty of the estimate (one standard deviation). Notice that the uncertainty on the estimate is considerably larger in the valleys between the harmonic peaks. This reflects the lower SNR in these regions. The vector shown is frame 100 from Figure 2

A second observation is that in voiced speech, the signal power is concentrated in areas near the harmonics of the fundamental frequency, which show up as parallel ridges in

[1] Even if the interfering signal is another speaker, the harmonic structure of the two signals may differ at different times, and the long term pitch contour of the speakers may be exploited to separate the two sources [2].

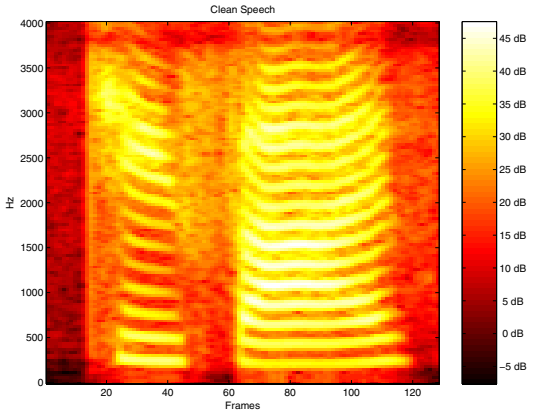
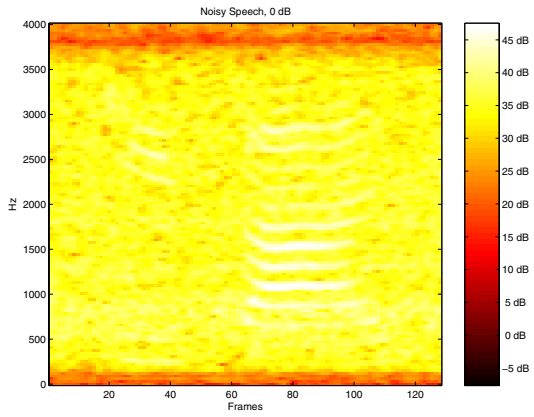


Fig. 2. Spectrogram of clean speech. The words 'TWO FIVE' are being spoken.

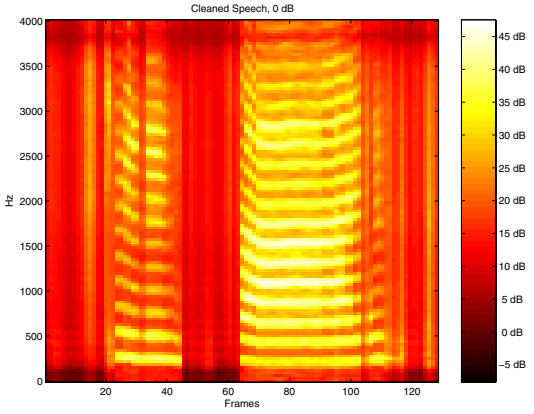
the spectrogram (see Figure 2). In a noisy environment, the local signal to noise ratio along the ridge is greater than the average SNR.

Figure 1 shows the estimate of a clean speech vector, the noisy input vector (car noise), and the true clean speech vector for comparison. The horizontal axis shows frequency in Hertz, and the vertical axis shows log-energy of the amplitude of each frequency. The regularly spaced peaks are the harmonics of the fundamental frequency. Notice that at the low end of the frequency range, the true signal is 'submerged' in the noise, whereas the harmonic peak at c.a. 670Hz and 900Hz emerge from the noise. Notice also that the first standard deviation (shown as a shaded area) of the estimate is large in the valleys, where the SNR is low and smaller around the harmonic peaks, where the SNR is higher. The method for producing the clean speech estimate is discussed in Section 2.

Researchers have sought to exploit this localization of signal power, both in the time domain and in the frequency domain. Methods for achieving this goal include alignment and gating of the glottal impulses in the time domain[3], and tracking the pitch as a pre-processing stage[4, 5]. Such approaches use highly constrained voicing models that are incongruous to the modelling of other aspects of the speech signal and employ modularized, multistage processing where aspects of the voicing are processed separately[6]. These approaches have been vulnerable to noise because of implicit independence assumptions or because the voicing estimation does not take noise into account. In addition, there may be excitation patterns and artifacts of the signal analysis that are poorly captured by such highly constrained models for harmonic structure. In contrast, our approach is to use a single high resolution log-spectrum model for both excitation and filter and train a model capable of capturing the relevant structures.



(a) Spectrogram of speech at 0 dB



(b) Spectrogram of cleaned speech at 0 dB.

2. MODEL BASED SIGNAL ENHANCEMENT

The core of the method involves calculating posteriors for the high frequency resolution log-spectrum $p(x|y)$, given the noisy speech. We employ the Algonquin framework [7, 8] to calculate these posteriors.

The model for noisy speech in the time domain is (omitting the channel for clarity)

$$y[t] = x[t] + n[t]. \tag{1}$$

where $x[t]$ denotes the clean signal, $n[t]$ denotes the noise, and $y[t]$ denotes the noisy signal. In the Fourier domain, the relationship becomes

$$Y(f) = X(f) + N(f) \tag{2}$$

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:

$$|Y(f)|\angle Y(f) = |X(f)|\angle X(f) + |N(f)|\angle N(f) \quad (3)$$

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

We model only the magnitude components and do not explicitly model the phase components. The relationship between the magnitudes is

$$|Y(f)|^2 = |X(f)|^2 + |N(f)|^2 + 2|X(f)||N(f)|cos(\theta) \quad (4)$$

where θ is the angle between X and N. For the purposes of modelling, we assume the we can model the last term as a noise term, hence we approximate this relationship between magnitudes as

$$|Y(f)|^2 = |X(f)|^2 + |N(f)|^2 + e \quad (5)$$

where the e is a random error [8]. Next we take the logarithm and arrive at the relationship in the high resolution log-magnitude-spectrum domain

$$y = x + \ln(1 + \exp(n - x)) + \varepsilon \quad (6)$$

where ε is assumed to be Gaussian. Hence, we can also write this relationship in terms of a distribution over the noisy speech features y as

$$p(y|x,n) = N(y; x + \ln(1 + \exp(n - x)), \psi) \quad (7)$$

where ψ is the variance of ε, and $N(y|\mu,\psi)$ denotes a normal density function in y with mean μ and variance ψ.

The transformations that we have applied to the model above are the same as the first steps in the calculation of the Mel frequency cepstrum features with the exception that we did not perform the Mel-scale warping before applying the log transform. For example, in the Aurora front end[9], the Mel-scale warping, smooths out the harmonics and reduces the dimensionality of the feature vector from 128 dimensions to 23 dimensions. The result of omitting the Mel-scale warping is that we do not smooth out the speech harmonics.

For the purpose of signal reconstruction, we are interested in likely values of clean speech, given the noisy speech. By recasting this relationship in terms of a likelihood $p(y|x,n)$, and using prior models for speech $p(x)$ and noise $p(n)$, we can arrive at a posterior distribution for the clean speech vector $p(x|y)$. This will be described in the next section.

By inverting the procedure described above we can reconstruct an estimate of the clean signal. To do this we find the MMSE estimate for clean speech $\hat{x}$ and calculate the inverse Fourier transform

$$\hat{x}[t] = IFFT(\exp(\hat{x}) \cdot \angle Y) \quad (8)$$

where $\hat{x} = \int x p(x|y) dx$. In this reconstruction, we have used the original phases from the noisy signal.

2.1. Inference

We now turn our attention to the procedure for estimating the posterior for the clean speech log-magnitudes $p(x|y)$. For this we employ the Algonquin method. Extensive evaluations of this framework have been performed in the context of robust speech recognition. In previous work, speech and noise models have either been in the "low-resolution" log-Mel-spectrum domain, or in the truncated cepstrum domain. Here we briefly outline the Algonquin procedure. Detailed discussions can be found in [7, 8].

At the heart of the Algonquin method is the approximation of the posterior $p(x|y)$ by a Gaussian.

The true posterior

$$p(x|y) = c \int p(y|x,n) p(n) p(x) dn \quad (9)$$

is non-Gaussian, due to the non-linear relationship in Eqn. (6). In Eqn. (9) c is a normalizing constant, $p(n)$ is the noise model, and $p(x)$ is the speech model, and $p(y|x,n)$ is the likelihood function discussed above.

We use a mixture of Gaussians to model both speech and noise. Hence

$$p(x) = \sum_s p(s) p(x|s) = \sum_s \pi_s N(x|\mu_s^x, \Sigma_s^x) \quad (10)$$

and similarly for $p(n)$. The construction of the speech model will be discussed below.

Due to the non-linear relationship between x and n for a given y, the true posterior $p(x|y)$ is non-Gaussian. We wish to approximate this posterior with a Gaussian posterior. The first step is to linearize the relationship between x and n.

For notational convenience, we write the stacked vector $z = [x^T n^T]$ and we introduce the function $g(z) = x + \ln(1 + \exp(n - x))$.

If we linearize the relationship of Eqn. (6) using a first order Taylor series expansion at the point z_0, we can write the linearized version of the likelihood

$$p_l(y|x,n) = p_l(y|z) = N(y; g(z_0) + G(z_0)(z - z_0), \Psi) \quad (11)$$

where z_0 is the linearization point and $G(z_0)$ is the derivative of g, evaluated at z_0. We can now write a Gaussian approximation to the posterior for a particular speech and noise combination as

$$p_l(x,n,y|s^x,s^n) = p_l(y|x,n) p(x|s^x) p(n|s^n) \quad (12)$$

It can be shown[8] that the $p(x,n|y,s^x,s^n)$ is jointly Gaussian with mean

$$\eta_s = \Phi_s \left[\Sigma_s^{-1}\mu_s + G^T \Psi^{-1}(y - g - Gz_0)\right] \quad (13)$$

and covariance matrix

$$\Phi_s = \left[\Sigma_s^{-1} + G^T \Psi^{-1} G\right]^{-1} \quad (14)$$

and the posterior mixture probability $p(y|s^x, s^n)$ can be shown to be

$$\gamma_s = |\Sigma_s|^{-1/2}|\Psi|^{-1/2}|\Phi_s|^{1/2} \cdot \exp\Big[-\frac{1}{2}(\mu_s^T \Sigma_s^{-1}\mu_s + (y_{obs} - g + Gz_0)^T \Psi^{-1}(y_{obs} - g + Gz_0) - \eta_s^T \Phi_s^{-1}\eta_s)\Big]. \quad (15)$$

The choice of the linearization point is critical to the accuracy of the approximation. Ideally, we would like to linearize at the mode of the true posterior. In the Algonquin algorithm, we attempt to iteratively move the linearization points towards the mode of the true posterior. In iteration i of the algorithm, the mode of the approximate posterior in iteration $i-1$, μ_{i-1} is used as a linearization point of the likelihood, i.e. $z_i = \mu_{i-1}$. The algorithm converges in 3-4 iterations.

2.2. Speech Model

Speech modelling for enhancement and speech recognition usually involves dimensionality reduction which removes the voice harmonics. This is either done explicitly, such as by using the Mel-warping, or implicitly, such as by using a small auto-regressive model. The filter and excitation components of the generative speech model are relatively independent, since voiced speech sounds can be spoken at any pitch. To model a particular speech sound in high resolution, one would expect to need an instance of the voiced acoustic model at each possible pitch.

A first approximation is to model the filter and excitation components independently. To construct such a model, one would lifter the 128 frequency component speech vectors to produce 128 component filter (vocal tract) features and 128 component excitation (vocal cords) features. This approach has the advantage that the models are compact, and independent temporal dynamics can be efficiently employed on each component, as in [2]. However, the model over-generates speech by allowing combinations of unvoiced excitation and voiced filters and vice versa, and the computations required for temporal dynamics may be too costly in many cases.

An alternate strategy is to simply train a single non-factored high resolution speech model. In the experiments described below, we used non-factored Gaussian mixture models (GMM). We trained two models: a speaker independent gender independent model, and a speaker independent gender dependent model. The speaker independent, gender independent model had 512 mixtures, and 128 frequency components, while the gender dependent model had 512 mixtures for the male component and 512 mixtures for the female component. These models were trained in the standard way[10], by initializing using vector quantization, and then using Expectation Maximization to find the parameters of the GMMs.

Although this approach is not as efficient as the factored model, with respect to the number of parameters required to represent combinations of voiced filters at different pitches, it has the advantage that it does not over-generate speech.

2.3. High Resolution Signal Reconstruction

To reconstruct the signal, we first calculate high resolution log-spectral features of the noisy input signal as described in Section 2. In the feature extraction stage, we used hamming windows of length 25 ms, and the frame rate of 10 ms. A corresponding synthesis window is designed such that the analysis window multiplied by the synthesis window, and overlapped with neighboring analysis-synthesis windows at the frame rate, sums to unity at each time point.

We smooth the high resolution log-spectrum features across frames by filtering them temporally with a simple FIR filter with parameters [0.25 0.5 0.25]. Without this smoothing step, the inference algorithm tends to produce spurious errors.

The Algonquin algorithm is then used to infer the posterior distributions over the clean speech. In the results reported below, we used the MMSE estimate based on $p(x|y)$. This is then exponentiated and used as a point estimate for $|X(f)|$. Alternately, we could use the MMSE estimate of $|\widehat{X(f)}|^2 = E[\exp(x)]$. However, the fact that the speech recognizer operates on the log spectrum domain motivates the former rather than the latter estimate.

We then reconstruct each frame of the signal, by use of the inverse Fourier transform, as in Eqn. (8), where the phase components are the phases of the noisy signal. The frames are then overlapped and added together using the tapered synthesis window described above.

3. RESULTS

We tested the high resolution signal enhancement for speech enhancement as well as for robust speech recognition.

3.1. Speech Enhancement Results

In informal listening tests, the subjective quality of the enhanced speech was reported to be exceptionally good. At very low SNR (-5 dB and 0 dB), the most notable distortion in the enhanced speech is flutter due to the inference algorithm assigning low energy fricatives to periods of silence, as well as silences in low energy voiced portions. At higher decibel levels (15 dB and 20 dB) the enhanced speech is almost indistinguishable from clean speech.

In Table 1 we give dB gains for the car noise condition of the Aurora data set. The first row shows SNR computed

	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
ΔSNR	10.76	8.38	6.27	3.95	1.28	-1.94
ΔSNR$_{seg}$	6.82	6.58	6.12	5.35	4.29	2.87

Table 1. Gains in Signal to Noise Ratio for Car noise at a range of SNR. The two measures of SNR are for standard SNR and Segmental SNR. For segmental SNR, we used a window of 25 ms, a SNR floor of -10 dB and an SNR ceiling of 35 dB.

over the whole waveform, while the second row shows segmental SNR, computed using a floor of -10 dB and a ceiling of 35 dB.

3.2. Aurora Speech Recognition Results

To assess the performance of high resolution signal reconstruction for speech recognition, we ran experiments on the Aurora 2 data-set. The Aurora 2 data-set contains spoken digits, artificially mixed with various noise types at Signal to noise ratios of -5 dB to 20 dB, in addition to unaltered clean speech. There are 1001 test files in each condition, where each test file contains from 1 to 7 spoken digits. In the experiments below, we report results for the Car noise condition. This condition has relatively stationary noise which allows us to use a single Gaussian noise model, estimated from the first 20 frames of each file. Other conditions such as "Subway" require larger noise models to handle the non-stationary aspect. In previous work, it has been shown [8] that using low-resolution Algonquin with larger noise models, as well as adapting the noise model will produce considerable gains in recognition accuracy, at the expense of higher computational complexity.

The standard low-resolution Algonquin method produces estimates of clean parameters in the 23 dimensional log-Mel-spectrum domain. For the recognition experiments, these are converted to cepstrum parameters directly, by taking the discrete cosine transform. For the high resolution signal reconstruction experiments, the time domain signal was reconstructed and the standard Aurora front end was then used to produce cepstrum parameters from the time domain signals.

The graph in Figure 3 shows the recognition accuracy for the Car noise condition of Set A of the Aurora 2 data-set, using multi-condition training of the acoustic models. We used the standard Aurora back-end, which is an HTK based recognizer with 16 state, left-to-right word models with 3 mixture acoustic models in each state. Figure 5 shows the change in absolute Word accuracy over the baseline, and Figure 4 shows the change in word error rate due to high resolution processing.

The baseline of 86.52% is shown as the bottom line in Figure 3. The result for "low-resolution" log-Mel-spectrum is the middle line in Figure 3. The speech model used was a Gaussian mixture model with 256 components, of 23 dimensions each. The low-resolution Algonquin algorithm achieves an average recognition accuracy of 90.12% for the Car noise condition, which is a relative reduction error rate of 13.26%.

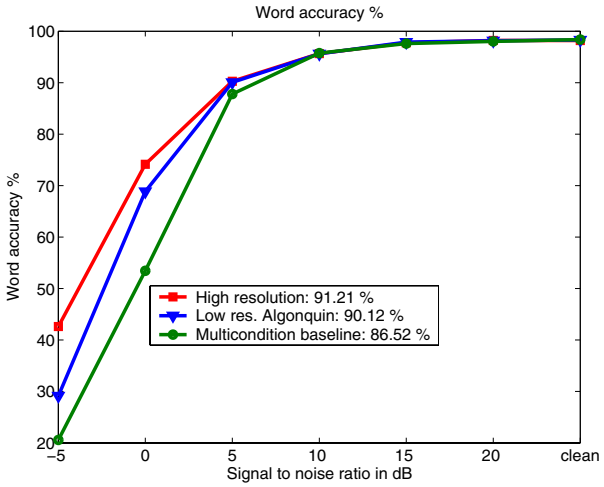


Fig. 3. Word Accuracy of High-Resolution Signal Reconstruction using Gender Dependent models, Low-Resolution Algonquin and Aurora Multicondition Baseline for the Car noise condition

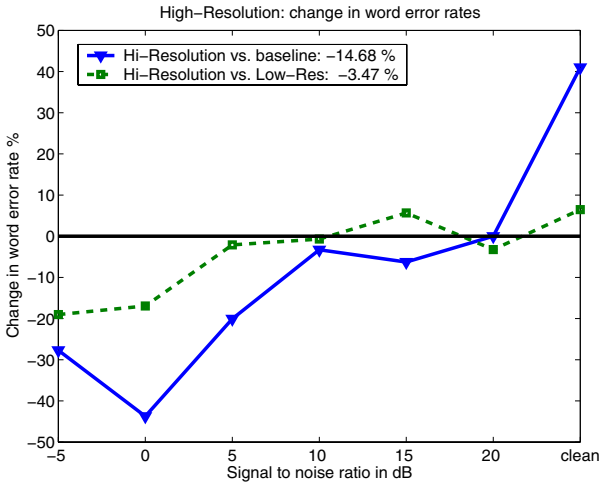


Fig. 4. Word error rate of High-Resolution method as compared to baseline, and Low-Resolution Algonquin.

The results for high resolution signal reconstruction with a speaker independent, gender dependent model is the top line in Figure 3. The average accuracy is 91.14%, which is a relative reduction in average word error rate of 15.62% over the baseline. Using gender independent high-resolution mod-

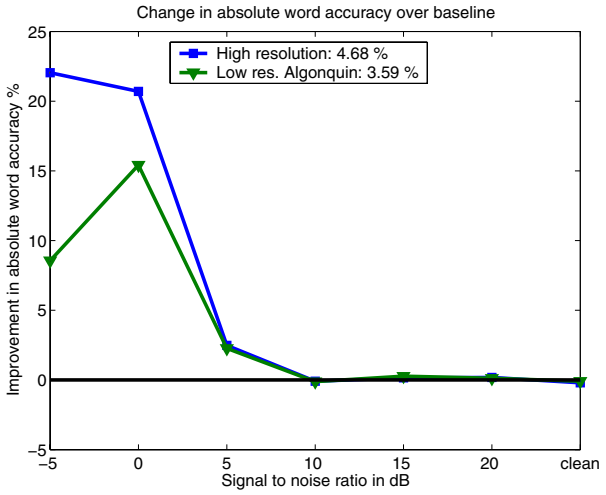


Fig. 5. Change in absolute Word Accuracy of High-Resolution Signal Reconstruction using Gender Dependent models and Low-Resolution Algonquin compared to the Aurora Multi-condition Baseline for the Car noise condition.

els achieves a slightly lower average accuracy of 91.04%.

It is more interesting to compare the recognition rates for low-resolution Algonquin and high-resolution Algonquin. Interestingly, the gains are mostly achieved at -5 dB and 0 dB. The increases in word accuracy are 5.28% and 13.48% absolute (16.95% and 19.02% reduction in WER respectively), while at higher SNRs the recognition rates are almost identical. This indicates that the advantages of using voicing information are mostly at very low signal-to-noise ratios. It also supports the assumption that voicing information is not helpful for speaker-independent recognition of clean speech in non-tonal languages.

4. DISCUSSION AND CONCLUSIONS

Our findings support the hypothesis that high resolution spectral information is quite useful for enhancing noisy speech and substantially helps recognition in very noisy conditions. At the same time, our findings are consistent with the widely held assumption that low-resolution spectral components are sufficient for speaker-independent recognition of clean speech.

The traditional approach for exploiting harmonic structure is to employ parametric models with a small number of parameters for the excitation component of the signal. This can lead to heterogeneous models and make it difficult to jointly estimate parameters related to excitation and filter in noisy conditions. The model presented in this paper avoids such pitfalls by employing a combined excitation-filter speech model. The size of model required is surprisingly small. Our model presents an advantage over models that factorize the excitation and filter components in that we can model statistical dependencies between the excitation and filter components of a signal.

We have incorporated this information into a probabilistic model in a principled way that is compatible with the current paradigm in speech processing.

5. REFERENCES

[1] Lawrence R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[2] J. Hershey and M. Casey, "Audio-visual sound separation via hidden markov models," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., Cambridge, MA, 2002, pp. 1173–1180, MIT Press.

[3] Dusan Macho and Yan Ming Chen, "Snr-dependent waveform processing for improving the robustnes of asr front-end," *In Proc. of ICASSP*, 2001.

[4] M. Seltzer, J. Droppo, and A. Acero, "A harmonic-model-based front end for robust speech recognition," *Eurospeech*, 2003, To appear.

[5] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Speech enhancement by harmonic modelling via map pitch tracking," *In Proc. of ICASSP*, pp. 549–552, 2002.

[6] S. Oberle and A. Kaelin, "Hmm-based speech enhancement using pitch period information in voiced speech segments," *International Symposium on Circuits and Systems ISCAS*, vol. 27, pp. 114–120, 1997.

[7] B.J. Frey, T. Kristjansson, L. Deng, and A. Acero, "Learning dynamic noise models from noisy speech for robust speech recognition," *Advances in Neural Information Processing (NIPS)*, 2001.

[8] T. Kristjansson, *Speech Recognition in Adverse Environments: A Probabilistic Approach*, Ph.D. thesis, University of Waterloo, Waterloo, Ontario, Canada, April 2002.

[9] Hans-Gunter Hirsch and David Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. of ISCA ITRW Workshop on Automatic Speech Recognition*, 2000.

[10] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the em algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, 1984.

[11] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 2, pp. 1526 – 1554, October 1992.