A HARMONIC-MODEL-BASED FRONT END FOR ROBUST SPEECH RECOGNITION

Michael L. Seltzer

Department of ECE
Carnegie Mellon University
Pittsburgh, PA 15213 USA
mseltzer@cs.cmu.edu

Jasha Droppo, Alex Acero

Speech Technology Group
Microsoft Research
Redmond, WA 98052 USA
{jdroppo,alexac}@microsoft.com

ABSTRACT

Speech recognition accuracy degrades significantly when the speech has been corrupted by noise, especially when the system has been trained on clean speech. Many robust techniques have been developed which require reliable online noise estimates or *a priori* knowledge of the noise. In situations where such estimates or knowledge is difficult to obtain, these methods fail. We present a new robustness algorithm which avoids these problems by making no assumptions about the corrupting noise. Instead, we exploit properties inherent to the speech signal itself to denoise the recognition features. In this method, speech is decomposed into harmonic and noise-like components, which are then processed independently and recombined. By processing noise-corrupted speech in this manner we are able to achieve significant improvements in recognition accuracy on the Aurora 2 task.

1. INTRODUCTION

The performance of automatic speech recognition systems degrades significantly when the speech signal is corrupted by additive noise. This is a major obstacle to the widespread deployment of speech recognition systems in real-world applications. Many algorithms have been proposed in the literature to compensate for the detrimental effect additive noise has on recognition performance. Many of these methods, such as [1], rely on an accurate estimation of the corrupting noise signal. This in itself is a very difficult problem in situations where the environmental noise is non-stationary. In such conditions, these methods fail.

Other methods rely on the use of noise models for compensation (*e.g.* [2]), or train a recognition system on noise-corrupted speech. When the test conditions are well-matched to the noise model or the training conditions, such methods perform well. However, it is impossible to account for the sheer variety of noises found in real world environments.

However, in all of these situations, the human user remains constant. That is, environmental noise has no affect on the speech production mechanism. As a result, we can improve the robustness of speech recognition systems by finding ways to exploit properties of the human speech signal itself. Algorithms created based on this premise can potentially perform well without making any assumptions about the noise signal or its properties.

One such feature of speech is the strong presence of a fundamental frequency and its harmonics in voiced speech. The fact that voiced speech has a well-understood, predictable harmonic structure makes this feature attractive as the basis for noise compensation algorithms. Several researchers have explored the use of the harmonicity of voiced speech for robust speech recognition. For example, Gu and Rose developed Perceptual Harmonic Cepstral Coefficients, which utilize a peak-picking algorithm to emphasize the harmonic spectral peaks in voiced speech [3]. In [4], Ealey *et al.* developed a harmonic "tunnelling" algorithm in which noise estimation is performed based on the nulls between the harmonic peaks in the spectrum which is used for spectral subtraction.

Over the last several years, the field of speech coding has benefitted from exploiting this harmonic structure of speech signals. Harmonic coding schemes are based on the principle that speech can be decomposed into a deterministic (also called periodic or harmonic) component and a noise-like or random component. Each of these signals can then be parameterized separately by exploiting the properties inherent in each one. Various researchers have proposed methods of performing such a decomposition. In [5], Yegnanarayana *et al.* use an iterative comb-filtering approach to perform the decomposition, while Laroche *et al.* proposed a harmonic+noise model in which a sum-of-sinusoids model is fit to the speech signal [6].

In this paper, we present an new algorithm for generating noise-robust feautures for speech recognition based on the harmonic+noise model (HNM) in [6]. Like most speech coding methods, the HNM aims to find a parameterization which most accurately represents the input signal. It has no noise-reduction capability, and does not differentiate between speech and environmental noise. As a result, the HNM will not inherently improve recognition accuracy.

However, the capability to decompose the speech signal into two different signals with known properties provides an appealing framework for noise compensation; once the signal has been split into its harmonic and random parts, each one can be processed independently, and then recombined to generate an enhanced signal. We present an novel extension of this model, called the *weighted harmonic+noise model*, and describe how it can be used to extract cleaner speech features from noise-corrupted speech in order to achieve significant improvements in recognition accuracy.

In Section 2 we review the harmonic+noise model, and discuss its application to noisy speech. In Section 3 we describe the proposed weighted HNM for improving the harmonic/stochastic decomposition in noisy speech in order to generate enhanced features. Experimental results evalulating our method are presented in Section 4. Finally, we summarize our findings in Section 5.

2. HARMONIC+NOISE MODEL OF SPEECH

The harmonic+noise model (HNM) of speech is based on the premise that a speech signal x is composed of a deterministic signal x_h and

a random signal x_r. It is assumed that the deterministic component is well-modeled as a sum of harmonically-related sinusoids given by

$$x_h(t) = \sum_{k=1}^{K} a_k \cos(k\omega_0 t) + b_k \sin(k\omega_0 t) \quad (1)$$

where ω_0 is the fundamental frequency and K is the total number of harmonics in the signal. Given a frame of speech, we would like to estimate the parameters of this harmonic model, namely the pitch or fundamental frequency ω_0 and the values of the amplitude parameters $\{a_1, a_2, \ldots, a_K, b_1, b_2, \ldots, b_K\}$. The pitch can be estimated using any number of pitch tracking algorithms in the literature. Given an estimate of the pitch, we can determine a least-squares solution for the amplitude parameters. To do so, we rewrite (1) in vector form as

$$\mathbf{x} = \mathbf{A}\mathbf{b}$$

where $\mathbf{x}$ is a vector of N samples, $\mathbf{A}$ is an $N \times 2K$ matrix given by

$$\mathbf{A} = [\mathbf{A_{cos}}\ \mathbf{A_{sin}}]$$

with elements

$$\mathbf{A_{cos}}(k,t) = cos(k\omega_0 t) \qquad \mathbf{A_{sin}}(k,t) = sin(k\omega_0 t)$$

and $\mathbf{b}$ is a $2K \times 1$ vector given by

$$\mathbf{b^T} = [a_1\ a_2 \ldots a_K\ b_1\ b_2 \ldots b_K]$$

Then, the least-squares solution for the amplitude coefficients is

$$\hat{\mathbf{b}} = (\mathbf{A^T A})^{-1} \mathbf{A^T x} \quad (2)$$

Using $\hat{\mathbf{b}}$, we can get an estimate for the deterministic portion of the speech signal, $\mathbf{x_h}$

$$\hat{\mathbf{x}}_\mathbf{h} = \mathbf{A}\hat{\mathbf{b}} \quad (3)$$

An estimate of the random component is then obtained simply as

$$\hat{\mathbf{x}}_\mathbf{r} = \mathbf{x} - \hat{\mathbf{x}}_\mathbf{h} \quad (4)$$

The HNM algorithm has no noise-reduction capability. It was designed to accurately capture the salient information present in the signal. Thus, when the HNM is applied to a speech signal corrupted by additive noise, the resulting harmonic and random components will be distorted by the noise. More explicitly, a HNM decomposition of noisy speech produces

$$\mathbf{y} = \mathbf{y_h} + \mathbf{y_r} \quad (5)$$

$$= \mathbf{x_h} + \mathbf{n_h} + \mathbf{x_r} + \mathbf{n_r} \quad (6)$$

where $\mathbf{n_h}$ is the portion of the noise signal which resides at the harmonics of the fundamental frequency and $\mathbf{n_r}$ is the noise at the non-harmonic frequencies. Thus, while a particular frame may have a given signal-to-noise ratio (SNR), the SNR of the harmonic and random components may be quite different depending on the energies of the speech and noise captured by each component.

If we have knowledge of the pitch and voicing state of the speech, we can use the harmonic model to help separate the signal from the noise. For example, in highly voiced frames, we know that a clean speech signal will be captured almost entirely by the harmonic component. Therefore, we can infer that any residual signal captured by the random component is mostly noise.

In the next section, we describe how we can apply the HNM to speech corrupted by additive noise to generate enhanced features for speech recognition.

3. A WEIGHTED HNM FRONT END FOR SPEECH RECOGNITION

Conventional Mel-frequency cepstral coefficients (MFCC) are derived for a frame of speech as follows. First a window is applied to a frame of speech, followed by a DFT. The power spectrum of the signal is then computed and the spectrum is smoothed using a series of triangular weighting functions applied along the Mel scale to capture the energy in a series of overlapping frequency bands. If we define $\mathbf{X}$ as the Mel spectral vector for a frame of speech $\mathbf{x}$, we can explicitly express the Mel spectrum as

$$\mathbf{X} = \mathbf{M}|DFT(\mathbf{x})|^2$$

where $\mathbf{M}$ is the matrix of Mel weighting coefficients. Finally, a truncated DCT is the applied to the logarithm of this Mel spectrum.

If we assume the harmonic and random components generated by the HNM are uncorrelated, the Mel spectrum of a frame of speech is simply the sum of the Mel spectra of the harmonic and random components. If we assume that the noise and the speech are also uncorrelated, then we can translate equations (5) and (6) directly into the Mel-spectral domain. Moreover, because Mel spectrum is a measure of energy, we can conclude that the observed noisy Mel spectral value is an upper bound on the actual clean speech value, *i.e.* $X \leq Y$, where X and Y represent a Mel spectral component of the clean and noise-corrupted speech, respectively.

Based on these observations, we can derive an estimate for the clean Mel spectral component of a noise-corrupted frame of speech within the framework of the HNM.

$$\hat{X} = \alpha_h Y_h + \alpha_r Y_r \qquad 0 \leq \alpha_h, \alpha_r \leq 1 \quad (7)$$

where Y_h, and Y_r are Mel spectral components of the harmonic and random signals of the observed speech frame, respectively, and α_h and α_r are scaling factors applied to these components. We call this the *Weighted Harmonic+Noise Model* (WHNM) to emphasize the fact that we are using scaled versions of the harmonic and random components to obtain an estimate of features of the underlying clean speech signal.

Clearly, the key to the success of this model is the accurate estimation of the scaling parameters α_h and α_r. It is apparent from (6) that they are a function of SNR. There are several potential methods for estimating these parameters. In this work, we chose to utilize the HNM framework itself in order to estimate these parameters. We assume for simplicity that a single scaling parameter can be applied to the entire Mel spectral vector. As mentioned in Section 2, the harmonic component of the HNM decomposition to capture most of the signal energy for strongly voiced frames when the signal is clean. As noise corrupts these frames, more energy will be present in the random component, and the proportion of the total signal energy captured by the harmonic component of the signal will decrease. This observation leads to an estimate for α_h.

$$\alpha_h = \frac{\sum_i y_h(i)^2}{\sum_i y(i)^2} \quad (8)$$

where the numerator represents the harmonic energy in the frame, and the denominator is the total energy of the frame. From (5), it is clear that this estimate will always be between 0 and 1. Figure 1 shows α_h for the frames of the utterance "2-7-oh-6-5-7-1" for clean speech and speech corrupted by subway noise to various SNRs from 20 dB down to -5 dB. The estimated voiced/unvoiced

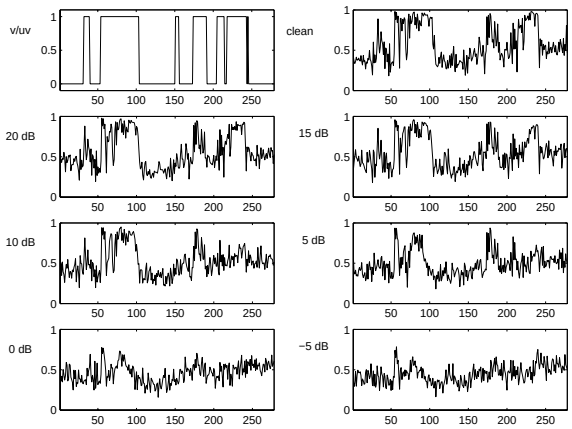

Fig. 1. α_h vs. time for an utterance corrupted by subway noise to various SNRs. The top left plot shows the voiced/unvoiced labelling.

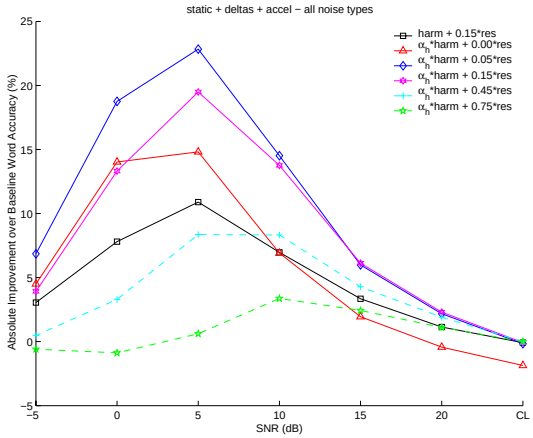

Fig. 2. Absolute improvement over baseline vs. SNR using various constant values of α_r using pitch estimates from clean speech.

labeling is plotted as well. The plots clearly show that α_h has a high correlation to the SNR in the voiced regions. In the unvoiced and silence frames, this measure of α_h serves simply as an energy-reduction parameter. While this estimate of α_h is sub-optimal for these segments, we found a significant benefit from processing all frames in the same manner, regardless of voicing state. This ensures that transition frames between voiced and unvoiced segments, whose harmonic and random components both contain significant information, are processed consistently. This reduces the frame-to-frame variability of the resulting features which is critical for accurate estimation of the delta and acceleration cepstal features used for recognition.

Obtaining an estimate for the scaling parameter of the random signal component is a more difficult task. Due to the very nature of the signal, there is no predictable underlying structure we can exploit. As with α_h, we expect α_r to be a function of SNR. In an attempt to learn this function from data, an experiment was performed studying the recognition performance obtained when a range of values of α_r are used with the estimate of α_h given by (8). Figure 2 shows the absolute improvement over baseline performance (no compensation) in word accuracy as a function of SNR for various values of α_r. In this experiment, the pitch estimation was performed on clean speech. As the plot indicates, there is a single value for α_r which results in the best performance across all SNRs.

Based on these results, we can rewrite the WHNM formulation for Mel spectral estimation as

$$\hat{\mathbf{X}}(t) = \alpha_h(t)\mathbf{Y_h}(t) + \alpha_r\mathbf{Y_r}(t) \qquad (9)$$

where the time index t has been added to emphasize that α_h is a time-varying parameter, while α_r is fixed. The value of α_h is computed according to (8) and α_r can be optimized using a cross-validation set.

4. EXPERIMENTAL RESULTS

To test the performance of the proposed Weighted Harmonic+Noise Model algorithm, experiments were conducted using the Aurora 2 corpus [7]. This corpus consists of strings of connected digits, corrupted by several different noises to SNRs between -5 and 20 dB. Speech recognition was performed using the HTK recognition system. The system was trained using the Aurora clean training set. Whole word models were trained using conventional 39 dimensional features vectors composed of 13-dimensional Mel-frequency cepstral coefficients plus the delta and accleleration features. In our experiments, a frame size of 20ms was used, rather than the 25ms specified in the Aurora standard, in order to reduce the variability of pitch within a frame. All other front-end, training, and testing specifications matched the Aurora specification.

The WHNM-based feature extraction process is as follows. For each utterance, pitch and voicing state estimation is performed. Frames labeled as non-voiced are assigned a pitch of 150 Hz. For each frame, the harmonic and random components of the signal are then computed using equations (2), (3), and (4). The Mel spectra of both the harmonic and random signals are computed and then the final Mel spectrum of the frame is computed as the weighted sum of the two, according to (9). Finally, the MFCC feature vector is computed by taking the DCT of the logarithm of the Mel spectrum.

As described in Section 3, the algorithm requires the use of a cross-validation set to determine the optimal value of α_r, the scaling parameter for the random component of the observed noisy Mel spectra. We employed the Aurora Test Set A for this purpose. The data set consists of utterances corrupted with one of four noises (babble, subway, car, exhibition hall) to SNRs between -5 and 20 dB.

For the cross-validation set, pitch estimates were made directly on the noise-corrupted speech data using the MAP pitch estimation method described in [8]. For each utterance, α_h was computed for each frame using (8). α_r was held constant over all utterances and a range of values were tested. Figure 3 shows the absolute improvement in word accuracy over baseline performance for the cross-validation test set for various values of α_r. For comparison, the performance achieved with the best-performing α_r value when using pitch estimates from clean speech is also shown.

As the figure indicates, setting α_r to 0.10 produced the best overall recognition accuracy. However, the figure also indicates that performance is not that sensitive to the value of α_r. It is in-

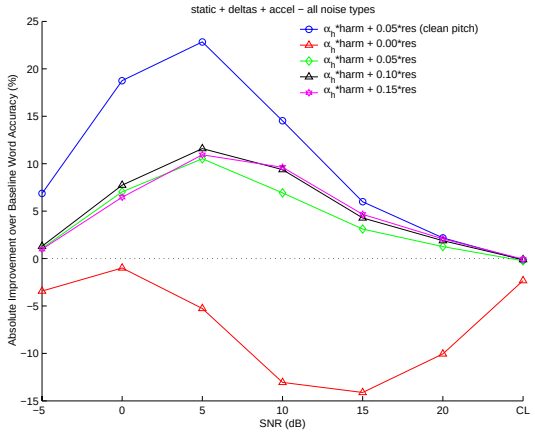

Fig. 3. Absolute improvement over baseline vs. SNR using various constant values of α_r on the cross-validation set when pitch estimates are made from noisy speech data. The performance using pitch estimates from clean speech is shown for comparion.

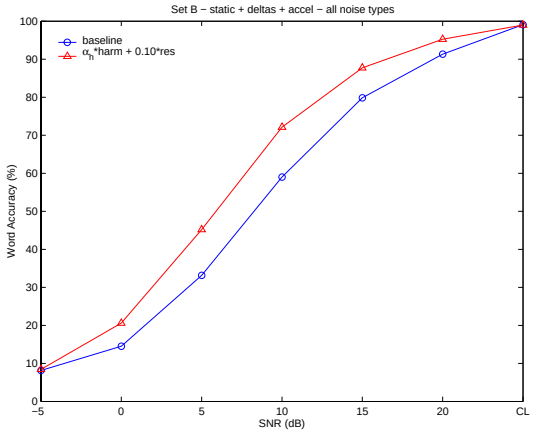

Fig. 4. % Word Accuracy vs. SNR for Aurora Test Set B using $\alpha_r = 0.10$ and pitch estimates from noisy speech.

teresting to note the extremely poor performance when α_r is set to 0 and the random component is completely removed from the signal. When this occurs, strict harmonicity is imposed on all segments of speech, including partially voiced and unvoiced frames. Because unvoiced and partially voiced speech contain discriminative information at the non-harmonic frequencies, removing this information results in poor performance.

Using the value of $\alpha_r = 0.10$ determined from cross-validation, recognition experiments were run on Aurora Test Set B. This test set consists of connected digit strings corrupted by four different noises (restaurant, street, airport, train station) to SNRs between -5 and 20 dB. There is no overlap between the cross-validation set and the test set. Figure 4 shows the recognition accuracy as a function of SNR when the proposed algorithm is applied to the Test Set B with $\alpha_r = 0.10$. The rightmost datapoint on the plot indicates the recognition accuracy on clean speech. For comparison, baseline performance without compensation is also shown. As the plot indicates, significant improvements over baseline recognition accuracy were achieved using the proposed method. A comparison of Figures 3 and 4 shows that the actual performance obtained on the test set was quite close to that obtained on the cross-validation set. From this we can conclude that the optimal choice of α_r is not sensitive to noise type, as the cross-validation and test sets had no overlap in corrupting noise types.

5. SUMMARY

The harmonic+noise model decomposes a speech signal into its harmonic and random components. This decomposition provides a framework in which these signal components can be processing independently, allowing us to exploit the properties inherent in each one. In this paper, we have used this framework to improve the robustness of speech recognition systems to additive noise. We introduced the weighted harmonic+noise model in the Mel-spectral domain in which the features are derived from the harmonic and random components independently, denoised, and then recombined to generate an enhanced final feature vector. By processing the noise-corrupted speech in this manner, we are able to achieve significant improvements in recognition accuracy without making any assumptions about the corrupting noise.

6. REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, April 1979.

[2] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, September 1996.

[3] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients for speech recognition in noisy environment," in *Proceedings of ICASSP*, Salt Lake City, Utah, May 2001.

[4] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunnelling: tracking non-stationary noises during speech," in *Proceedings of Eurospeech*, Aalborg, Denmark, September 2001.

[5] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 1–11, January 1998.

[6] J. Laroche, Y. Stylianou, and E. Moulines, "HNM: A simple efficient harmonic + noise model for speech," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, NY, October 1993.

[7] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millenium*, Paris, France, September 2000.

[8] J. Droppo and A. Acero, "Maximum a posteriori pitch tracking," in *Proceedings of ICSLP*, Sydney, Australia, December 1998.