
Improving the Front-End Noise Preprocessor of MELPe

Xin Lei, Mari Ostendorf

{leixin, mo}@ee.washington.edu

Dept of Electrical Engineering, University of Washington

Seattle WA, 98195-2500

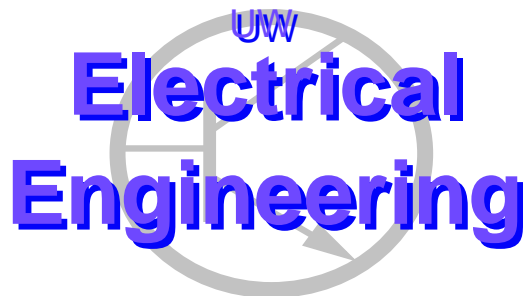
Lane Owsley

lane@apl.washington.edu

Applied Physics Laboratory, University of Washington

Seattle WA, 98105-6698

UWEE Technical Report
Number UWEETR-2004-0000
July 2004



Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: <http://www.ee.washington.edu>

Improving the Front-End Noise Preprocessor of MELPe

Xin Lei, Mari Ostendorf
{leixin,mo}@ee.washington.edu
Dept of Electrical Engineering, University of Washington
Seattle WA, 98195-2500

Lane Owsley
lane@apl.washington.edu
Applied Physics Laboratory, University of Washington
Seattle WA, 98105-6698

University of Washington, Dept. of EE, UWEETR-2004-0000

July 2004

Abstract

In this paper we focus on improving the noise preprocessor (NPP) of the low-rate speech coder MELPe using information from the non-acoustic General Electromagnetic Motion Sensor (GEMS). A generalized linear model approach is proposed to improve the voice activity estimation both in the frame-level time domain and in the bin-level frequency domain with GEMS and context features. HMM based speech recognition techniques are also investigated to drive the estimators. The improved voice activity parameter estimators are shown to have significantly less error than the estimates from MELPe NPP. The improved frame-level voice activity estimator achieves 66% reduction in error. The improved bin-level voice activity estimates has more than 50% error reduction. With an optimal spectral amplitude estimation algorithm instead of the MM-LSA algorithm used in MELPe NPP, and the improved voice activity parameters, the processed noisy speech has much less residue noise and higher intelligibility in informal listening tests.

1 Introduction

The enhanced Mixed Excitation Linear Prediction (MELPe) [1] vocoder is a dual-rate low-rate coder that operates at 1200 and 2400 bps. The Defense Advanced Research Projects Agency (DARPA) sponsored Advanced Speech Encoding (ASE) program aims to improve the low rate MELPe coder in noisy military environments using multiple non-acoustic sensors [2]. As the first step of this project, our efforts focused on sensor information integration to improve the front-end noise preprocessor (NPP) of MELPe.

The noise preprocessor of MELPe involves an initial stage of voice activity detection (VAD), and then uses the multiplicatively-modified log-spectral amplitude (MM-LSA) estimator for speech enhancement [3, 4]. This log-spectral amplitude (LSA) estimator uses a soft-decision gain modification to indicate the narrowband speech presence probability for each frequency bin [5]. Compared to the frequency bin-level soft-decision gain, the VAD provides a broadband indicator of whether there is speech in a particular frame. These two speech activity indicators are very important to the noise preprocessing: the broadband VAD allows for updating spectral estimates of the noise, and the bin-level indicator helps account for changing speech spectral characteristics in the noise filtering equation. Unfortunately, the estimates of these two terms from the MELPe noise preprocessor are not very accurate in severe noise environments, resulting in poor quality speech. Our pilot experiments showed that the inaccuracy in the estimation of these two terms is a critical problem in the MELPe noise preprocessor.

In this work, we introduce approaches to improve the broadband and narrow band speech activity estimates by introducing context information as well as features from non-acoustic sensors. The non-acoustic sensor investigated in this work is the General Electromagnetic Motion Sensor (GEMS) [6]. In addition, we investigate the use of speech

recognition techniques based on hidden Markov models (HMMs) to drive the estimators. Finally, we assess the interaction of these improvements with an optimally modified LSA (OM-LSA) estimator [7].

Section 2 will briefly review the mean-square error (MSE) based short-time LSA estimation based speech enhancement algorithms. Section 3 will describe the generalized linear model (GLM) based broadband (VAD) and narrowband (soft decision gain) speech activity detection using GEMS and context information. Section 4 will discuss the HMM-based approaches, applying speech recognition techniques to improve the speech enhancement process. The performance improvements of the two approaches are listed within the sections they are discussed. Section 5 will discuss the results of informal listening tests using the different approaches. Section 6 will summarize the key findings and discuss possible future work.

2 Review of LSA estimation

Assume $x(n)$ and $d(n)$ denote uncorrelated speech and additive noise respectively. We have the noisy signal $y(n) = x(n) + d(n)$. By applying short-time Fourier analysis on the overlapping frames, in frequency domain we have $Y_k = X_k + D_k$, where $X_k = A_k e^{j\phi_k}$ and $Y_k = R_k e^{j\theta_k}$, with k denoting the frequency bin index. The minimum mean-square error LSA (MMSE-LSA) estimator [9] is shown to have better performance than the minimum mean-square error short time spectral amplitude (MMSE-STSA) [8]. The MMSE-LSA estimator is given in [3]:

$$\begin{aligned}\hat{A}(k) &= \exp\{E[\ln A_k | Y_k, H_1^k] G_M(k)\} \\ &\triangleq [G_{LSA}(k) R_k]^{G_M(k)}\end{aligned}\quad (1)$$

where H_1^k indicates speech presence in k -th frequency bin, $G_M(k) = P(H_1^k | Y_k)$ is the soft-decision modification denoting the speech presence probability in k -th bin in the specified frame [5], and $G_{LSA}(k)$ is the gain function derived in [9].

The MELPe NPP adopts the multiplicatively-modified LSA (MM-LSA) estimator [3], which is a linearly approximated version of MMSE-LSA:

$$\hat{A}(k) = G_M(k) G_{LSA}(k) R_k \triangleq G_{MM-LSA} R_k \quad (2)$$

Since the MM-LSA is an approximation and not optimal, Cohen proposed an optimally modified LSA (OM-LSA) in [7]. By constraining the soft-decision gain G_M to be larger than a threshold G_{min} for speech absent bins, Cohen derived a spectral amplitude estimator that is nonlinear with the soft-decision gain G_M :

$$\begin{aligned}\hat{A}(k) &= \{G_{LSA}(k)\}^{G_M(k)} \{G_{min}(k)\}^{1-G_M(k)} R_k \\ &\triangleq G_{OM-LSA} R_k\end{aligned}\quad (3)$$

By applying the attenuation term $G_{MM-LSA}(k)$ or $G_{OM-LSA}(k)$ to each spectral component R_k , we can get the spectral estimator $\hat{A}(k)$ of the clean speech. In this work, we tried both alternatives to estimate the spectral amplitude. Since accurate phase is not essential in speech enhancement, by using the phase of the noisy speech as the clean speech phase, the clean speech can be obtained from an inverse Fourier transform.

3 Broadband and narrowband speech activity detection

Voice activity detection (VAD) is a broadband speech activity detector for a particular frame, while the soft-decision gain $G_M(k)$ is a narrow band speech activity detector for the k -th frequency bin. They are both essential to speech enhancement algorithm. However, when processing noisy speech, the MELPe NPP often gives very poor estimates of these two parameters. We conducted informal listening tests to assess whether the frequent speech activity detection errors led to degraded speech quality by adding noise to a clean signal and using hand-labeled VAD decisions in the MELPe algorithm. As we expected, there was a clear improvement in quality associated with the hand-labeled VAD.

Motivated by this and other oracle experiments with the soft-decision gain term, we investigated improved estimators for these terms. Since both VAD and G_M involve probabilities associated with a binary indicator of speech activity, both terms can be estimated in the same mathematical framework. We use a very simple model, described in Section 3.1, which integrates temporal and bandwidth context and additional sensor data. Quantitative analyses of the impact on performance in experiments with artificially added noise are presented in Section 3.2.

3.1 Generalized linear model

We propose a general approach to utilize the features with a generalized linear model (GLM). A generalized linear model has the following form:

$$f(Y) = \sum_{i=1}^d a_i x_i + b \quad (4)$$

where $f(\cdot)$ is called the link function, Y is the predicted value, x_i 's are the input features, a_i and b are constants. In our case, the VAD and each bin-dependent G_M are probability terms that correspond to binary variables, so we choose a *logit* link function and get the logistic regression model:

$$\ln\left(\frac{p}{1-p}\right) = \sum_{i=1}^d a_i x_i + b \quad (5)$$

Therefore, with a set of features $\{x_i : i = 1, \dots, d\}$ from context or other sensors, we can get an estimator of the speech activity probability variable p :

$$\hat{p} = \frac{\exp(\sum_{i=1}^d a_i x_i + b)}{1 + \exp(\sum_{i=1}^d a_i x_i + b)} \quad (6)$$

The parameters of the model, $\{a_i : i = 1, \dots, d\}$ and b are learned from data by using hand-labeled target values paired with the feature vector $x = [x_1, \dots, x_d]$. Details on features used for each of the two models are given next. All generalized linear models in this work are trained and tested using the S-Plus statistics package.

3.2 Feature selection

The VAD in MELPe NPP is based on the estimation of the *a posteriori* SNR [3] which is a very rough estimate especially in severe noise environments. GEMS is a novel sensor which measures the glottis movement and is relatively immune to the acoustic noise. By extracting the features from the GEMS, a much more stable VAD can be achieved. Also, since the speech and silence regions tend to appear in chunks, it is possible to further improve the VAD performance by combining temporal context information from the MELPe VAD and GEMS in the current and neighboring ± 3 frames. We explored different combinations of these features as the input x to the logistic regression model given by equation (6), including: 1) the VAD from the MELPe NPP and its context; 2) the MELPe VAD, its context and GEMS log energy for current frame (GEMS VAD); 3) GEMS VAD only; and 4) GEMS VAD and its context. Comparing the GLM output probability to a probability threshold 0.5, a decision can be made on the speech activity in the specified frame.

Considering speech is non-stationary and may have different presence probability characteristics in different frequency bins, we trained a soft-decision GLM for every frequency bin. The goal of the G_M GLM is to minimize the error between the G_M estimates of the noisy utterances and the reference values. Since G_M indicates the speech presence activity in a particular frequency bin in the specified frame, the context features of G_M should include both temporal and frequency contexts. The context windows span ± 3 frames in time and ± 3 bins in frequency. However, in this case we do not use explicit values but rather averages and first (and second) differences to reduce the number of parameters. We chose a size-8 G_M feature vector that contained both time context and frequency context information of the noisy G_M . A GLM was trained for each frequency bin except the boundary frames and the boundary bins. The output probability of the GLM can be used directly as the G_M estimate.

3.3 Experimental results

The DARPA ASE pilot corpus [2] was used for our experiments. The pilot corpus includes speech from 10 male and 10 female speakers in nine different environments. Six different sensors were recorded for each speech utterance. In our experiments we only used GEMS. We performed the experiments on the 20 consonant vowel consonant (CVC) utterances recorded in a quiet environment. Each of the CVC utterances is around 90 seconds. CVC utterances from all speakers except F4, F5, M4 and M5 are used for training. Testing is performed on utterances from F4, F5, M4 and M5. We added 10dB M2 Bradley Fighting Vehicle noise to the clean utterances, to use as a controlled set of the noisy utterances that had "known" (from the clean versions) target VAD values. Unfortunately, it appeared that even

for clean versions, the VAD estimates from MELPe were unreliable (though better than the noisy case), so we hand-labeled VAD targets for all of the training and testing data. This hand-labeled VAD was substituted into the MELPe algorithm to generate new G_M values from the clean data for use as targets for training and evaluation.

The accuracy results of the various VAD algorithms compared to the hand-labeled VAD on the testing set are listed in Table 1. The best result is from GLM combining GEMS and context features. This VAD accuracy corresponds to a 66% reduction in error and is obtained by using only the GEMS signal, i.e., ignoring the noisy speech signal entirely.

Table 1: VAD generalized linear model accuracy results.

Feature Set	VAD Accuracy (%)
MELPe VAD	74.31
MELPe VAD + context	76.55
MELPe VAD + context + GEMS VAD	86.32
GEMS VAD	88.50
GEMS VAD + context	91.14

The G_M GLMs were trained and tested on the same training and testing subsets as the VAD GLM. Unlike the VAD experiments, for the G_M case we used the continuous-valued probability provided by MELPe (given the hand-labeled VAD) as the target for both prediction and evaluation. Hence, the evaluation is in terms of root mean square error (RMSE) with respect to the G_M “truth” rather than frame accuracy. We found the G_M GLM generally gave a reduction in RMSE of more than 50% compared to the noisy G_M . Example frequency bin results are listed in Table 2, where the third column corresponds to the HMM-based extension described in the next section.

Table 2: G_M generalized linear model RMSE results.

Frequency Bin	Noisy G_M	G_M GLM	Augmented G_M GLM
11	0.445	0.232	0.231
67	0.411	0.202	0.199
120	0.389	0.183	0.183

4 HMM-based approaches

The current MELPe NPP is designed for noise processes that are only slowly varying, and represents only two signal states (speech and silence/noise) according to the VAD decision. The hidden Markov model (HMM) has been widely accepted by the speech community as a reliable statistical model for speech signals. As opposed to the continuous-state model effectively used in MELPe, the HMM models speech as coming from a system in one of a large number of discrete states. As a result, the HMM can capture longer-distance phonotactic patterns and increase reliability of the estimate of the speech state, and these can be combined with a continuous-state model [10]. Further, speech and noise HMMs can be combined in a product model to effectively handle non-stationary noise conditions. Finally, the HMM offers a simple framework for integrating information from other sensors in detecting speech states. For all these reasons, we chose to investigate the use of HMMs for improving noise preprocessing in MELPe. Due to time limitations, much of the work focused on improving HMM performance, but we also investigated using the HMM to drive the VAD and the G_M estimators.

4.1 Method

Since the goal was speech state detection and not word recognition, and since we did not want a large state space for complexity reasons, we chose to build an HMM based on simple context-independent phone models (also called monophones). There is very little data in the ASE corpus, so we first trained a speaker-independent monophone recognition system on TIMIT database, which has 4620 short training sentences. Three systems were then designed:

one on the original TIMIT data, one on the TIMIT data with added M2 Bradley Fighting Vehicle noise, and a TIMIT-trained system that was adapted to the ASE noisy CVC training utterances using maximum likelihood linear regression (MLLR) [11].

The acoustic features are the standard Mel-Frequency Cepstral Coefficients (MFCCs) plus the log energy and their deltas and second deltas, leading to a 39-dimensional feature vector. Our system has 48 context independent phoneme models. Each model has 3 emitting states except the silence model. Each state has 32 Gaussian mixture components. The 48 phones are mapped from the 64 TIMIT phones following the method presented in [12]. The phone HMMs are trained with the hidden Markov model toolkit (HTK) [13]. Because the HMM is evaluated on CVC data that does not reflect the same phonotactics in naturally spoken English, we used a simple unigram model to characterize phone sequence probabilities.

In the context of the MELPe NPP, we used the HMM-based phone recognition system to drive the estimation of the VAD and G_M . From the HMM Viterbi-decoded phone-level states, a straightforward VAD decision can be made based on whether the decoded phoneme is silence or not. This method can be seen as an alternative method for VAD estimation. To improve G_M estimation, there are several options for using the HMM. In the experiments reported here, we append the HMM-decoded phoneme state as an extra feature for the G_M GLM, which effectively adds a bias term. In order to reduce complexity, particularly because of limited training data, we collapse the phone states into 6 broad classes with roughly similar spectral characteristics: fricative, nasal, vowel and liquid, stop, silence, and stop closures.

4.2 Experimental results

The reference values for the ASE test data were obtained from forced alignments on all the clean CVC utterances, given a dictionary of phone-based expansions of each word. On ASE CVC test utterances, the broad class recognition accuracy of the noisy TIMIT trained HMM recognition system is 47.55%. The MLLR adapted HMM's improve the accuracy to 56.65%.

For reference, the clean TIMIT trained monophone recognition system gives 65.1% phone accuracy¹ on the clean TIMIT test set, which is comparable to other reported monophone results. The noisy TIMIT trained monophone recognition system has 45.4% phone accuracy on the noisy TIMIT test set.

The first application of the HMM decoded results is to get a silence detector and use it as VAD. We tested this HMM driven VAD on utterance from speaker F4, it gives VAD accuracy of 82.6%, while the GEMS and context based GLM VAD gives accuracy of 84.04%. Therefore, the HMM-based VAD is comparable to the GEMS-based GLM VAD in this case. The HMM-based VAD has the possible advantage of less missed speech than GEMS-based VAD since GEMS cannot distinguish silence and unvoiced speech. An HMM-driven G_M GLM is the second application. We appended the HMM decoded broad class labels as a categorical feature to the G_M GLM feature vector and retrained the models in S-Plus. On the testing set we find that the augmented G_M GLM's give a lower mean-square error for most of the frequency bins, though the improvement is not significant. Results of a few bins are listed in Table 2.

5 Informal listening experiments

Using the GEMS-based VAD estimators, HMM-driven VAD and the augmented G_M GLM, we processed the speech utterance from speaker F4. We also tried the OM-LSA algorithm in the speech processing, setting the threshold $G_{min} = 0.1$. Segments of the processed speech waveforms are shown in Figure 1.

In informal listening tests, it was found that several large periods of noise residue in the MELPe NPP-processed speech were attenuated with the improved VAD and G_M estimates, as seen in Figure 1 most clearly in the 21-22s and 24.5-26s time periods. The GLM VAD and HMM VAD processed results are very similar. The G_M GLMs produced speech that was more natural sounding, though the perceived noise level was not significantly affected. The OM-LSA algorithm resulted in speech with much less noise in the non-speech regions, but had a bit more distortion in the speech regions.

¹On TIMIT corpus we evaluate the results in terms of monophone accuracy to compare with other work.

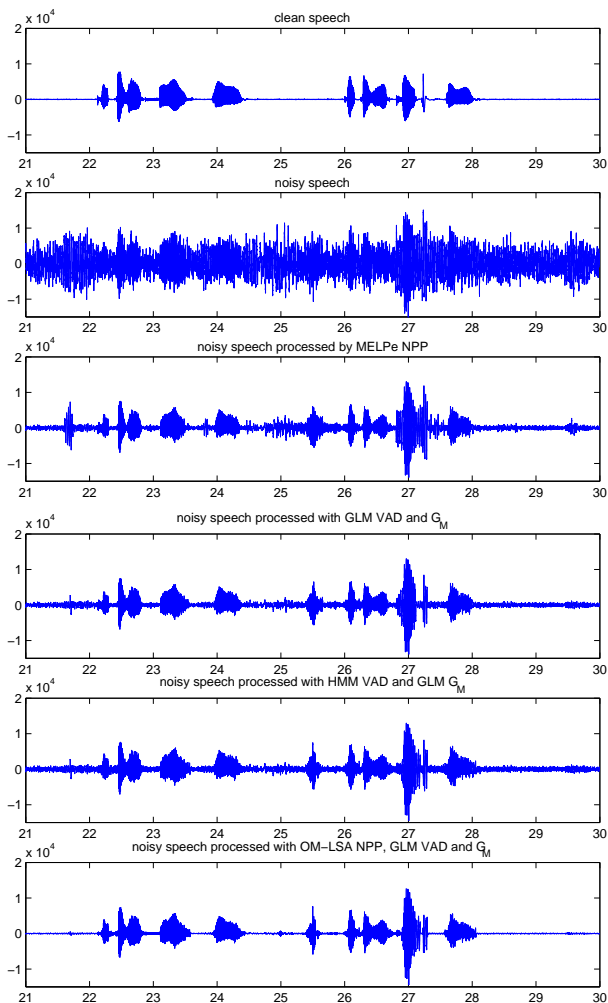


Figure 1: *Illustration of Noisy Speech Processed.*

6 Conclusions

The results reported have shown that by taking a generalized linear model approach, the VAD and soft decision gain G_M can be significantly improved by incorporating GEMS and context information. HMM state decoding provides useful information on its own for the VAD that might offer further improvements in combination with the GLM. Improvements to the soft decision gain term from the HMM state were not statistically significant, but other mechanisms for integrating this information remain to be explored.

7 Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency under Contract N00024-02-D-6602; TD 6101003. The opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

References

- [1] Wang, T., Koishida, K., Cuperman, V., Gersho, A. and Collura, J., "A 1200/2400 bps Coding Suite Based on MELP", 2002 IEEE Workshop on Speech Coding, Tsukuba, Japan, 6-9 October 2002.

- [2] Quatieri, T.F., Messing, D., Brady, K., Campbell, W.B., Campbell, J.P., Brandstein, M., Weistein, C.J., Tardelli, J.D. and Gatewood, P.D., "Exploiting Nonacoustic Sensors for Speech Enhancement", in Workshop on Multimodal User Authentication, 2003.
- [3] Malah, D., Cox, R.V. and Accardi A., "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments", Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP), 1999.
- [4] Martin, R. and Cox, R.V., "New Speech Enhancement Techniques for Low Bit Rate Speech Coding", Speech Coding Proceedings, pp. 165-167, IEEE Workshop, 1999.
- [5] McAulay, R. and Malpass, M., "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", IEEE Trans. Acoustics, Speech and Signal Processing, vol. 28, pp. 137-145, December 1980.
- [6] Ng, L.C., Burnett, G.C., Holzrichter, J.F. and Gable, T.J., "Denoising of Human Speech Using Combined Acoustic and EM Sensor Signal Processing", Proc. ICASSP, pp. 229-232, 2000.
- [7] Cohen, I., "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator", IEEE Signal Processing Letters, vol. 9, pp. 113-116, April 2002.
- [8] Ephraim, Y. and Malah, D., "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Trans. Acoustics, Speech and Signal Processing, vol. 32, pp. 1109-1121, December 1984.
- [9] Ephraim, Y. and Malah, D., "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator", IEEE Trans. Acoustics, Speech and Signal Processing, vol. 33, pp. 443-445, April 1985.
- [10] Ostendorf, M., Digalakis, V. and Kimball, O., "From HMMs To Segment Models: A Unified View of Stochastic Modeling For Speech Recognition", IEEE Trans. Acoustics, Speech and Signal Processing, vol. 4, pp. 360-378, April 1996.
- [11] Legetter, C. and Woodland, P., "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs", Computer Speech and Language, vol. 9, No. 2, pp. 171-185, 1995.
- [12] Lee, K.F. and Hon, H.W., "Speaker-Independent Phone Recognition Using Hidden Markov Models", IEEE Trans. Acoustics, Speech and Signal Processing, vol. 37, pp. 1641-1648, November 1989.
- [13] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V. and Woodland P., The HTK Book (for HTK Version 3.1), Microsoft Corporation and Cambridge University Engineering Dept, 2001.