Discriminative Training for Multiple Observation Likelihood Ratio Based Voice Activity Detection

Tao Yu, *Student Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—It is possible to show that the likelihood ratio (LR) test from multiple observations can enhance the performance of a statically modeled voice actively detection (VAD) system. However, the combination weights for the likelihood ratios (LRs) in each observation are rather empirical and heuristical. In this study, the optimal combination weights from two discriminative training methods are studied to directly improve VAD performance, in terms of reduced misclassification errors and improved receiver operating characteristics (ROC) curves. As shown in the evaluations, VAD performance, both in terms of absolute performance and consistency across noise types, can be significantly improved using the proposed method.**

***Index Terms*—Discriminative training, receiver operating characteristics (ROC), voice activity detection (VAD).**

I. Introduction

An important problem in statistically modeled voice actively detection (VAD) is that fluctuations in the instantaneous likelihood ratio (LR) generates a high *miss-hit rate* in the speech offset region or *false-alarm rate* in the noise nonstationary region. Therefore, VAD decisions are generally made from multiple observations rather than a single instantaneous observation, taking advantage of the strong correlation in the consecutive time-frames of speech. A "hangover" scheme based on hidden Markov model (HMM) was previously explored in [1]. Later, a simple but effective first-order smoothed LR scheme was employed in [2]. Recently, a multiple observation likelihood ratio test (MO-LRT) was considered in [3] and shown to outperform traditional decision smoothing methods.

While testing statistics from multiple observations could reduce detection errors, the strategy to effectively utilize them is rather empirical and heuristical. In [4], a discriminative training method is introduced in the statistically modeled VAD context, however, more advanced methods are possible. In the present study, two discriminative training methods are further studied for effective combination of multiple observation LRs, in terms of misclassification errors and receiver operating characteristics (ROC) curves, as shown in Section III. Next, an extensive set of evaluations is conducted in Section IV with conclusions drawn in Section V.

II. Problem Formulation

A. Signal Model and Single Observation LLR

Here, assume that the speech s is degraded by an uncorrelated additive noise n. Under two hypotheses H_0 (*speech-pause*) and H_1 (*speech-active*), the observation x in the short-time Fourier transform (STFT) domain can be written as

$$\begin{aligned} H_0(\mathrm{speech-pause}) &: x_{k,t} = n_{k,t}, \\ H_1(\mathrm{speech-active}) &: x_{k,t} = s_{k,t} + n_{k,t} \end{aligned} \tag{1}$$

where k and t are the frequency-bin and time-frame index, respectively. Suppose there is in total of K frequency-bins, and denote $\boldsymbol{x}_t = \{x_{1,t}, x_{2,t}, \ldots, x_{K,t}\}^T$ as a vector contains all the STFT coefficients in the tth time-frame, a statistic for a time-frame-wise VAD decision can be obtained from maximal *a posteriori* (MAP) criterion as

$$\begin{aligned} l_t &= \log(p(H_1|\boldsymbol{x}_t)) - \log(p(H_0|\boldsymbol{x}_t)), \\ &= \log\frac{p(H_1)}{p(H_0)} + \log(p(\boldsymbol{x}_t|H_1)) - \log(p(\boldsymbol{x}_t|H_0)) \end{aligned} \tag{2}$$

where l_t is the log-likelihood ratio (LLR) of the tth time-frame; $p(H_i)$ with $i \in \{0,1\}$ is the *a priori* probability of either *speech-pause* or *speech-active* and in principle does not depend on the observation; also, $p(\boldsymbol{x}_t|H_i)$ is the likelihood based on the probability density function (pdf) modeled for H_i [1], [5]. With this, the decision rule can be established as

$$l_t = \log p(\boldsymbol{x}_t|H_1) - \log p(\boldsymbol{x}_t|H_0) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{3}$$

where η is the detection threshold that controls the tradeoff between the miss-hit rate and false-alarm rate. Clearly, this decision rule is based on the instantaneous LLR, whose fluctuations lead to serious false-alarm errors and miss-hit errors as previously discussed.

B. Multiple Observation LLRs

A more sophisticated method is to incorporate contextual information into the decision rule. Suppose that a collection of M sequential LLRs from the current time-frame t, denoted as $\boldsymbol{l}_t = \{l_t, l_{t-1}, \ldots, l_{t-M+1}\}^T$, is used to make VAD decision for the current time-frame t, a new statistic that reflects the dependence on the current time-frame as well as its previous $M-1$ time-frames, can be expressed as

Manuscript received May 15, 2010; revised July 02, 2010; accepted July 05, 2010. Date of publication August 12, 2010; date of current version September 02, 2010. This work was supported by the Air Force Research Laboratory through a subcontract to RADC, Inc. under FA8750-09-C-0067, and in part by the University of Texas at Dallas, from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen. Approved for public release; distribution unlimited. The associate editor coordinating th review of this manuscript and approving it for publication was Dr. Constantine L. Kotropoulos.

The authors are with the Center for Robust Speech System (CRSS), the University of Texas at Dallas, Richardson TX 75080 USA (e-mail: tao.yu@student.utdallas.edu; john.hansen@utdallas.edu).

Color versions of one or more figures in this paper are available at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2010.2066561

$$\tilde{l}_t = w_1 l_t + w_2 l_{t-1} + \cdots + w_M l_{t-M+1} = \boldsymbol{w}^T \boldsymbol{l}_t \tag{4}$$

where $\boldsymbol{w} = \{w_1, w_2, \ldots, w_M\}^T$ is a vector of the combination weights for different time-frames. The decision rule can then be established as

$$\tilde{l}_t = \boldsymbol{w}^T \boldsymbol{l}_t \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{5}$$

The criterion on how to choose the combination weights w_k is the primary focus of this study. Under the assumption of independence between each time-frame, equal weighting was previously studied in [3] and showed better performance over conventional VADs. However, the optimality of equal weights may be challenged from two aspects: strong correlation exists between consecutive noisy speech frames and an unclear relation with the overall VAD performance. These issues are addressed using discriminative training in this study.

III. Discriminative Training

Suppose there is a set of labeled LLRs for training, denoted as $\mathcal{L} = \{\mathcal{L}^s, \mathcal{L}^p\}$, where $\mathcal{L}^s = \{\boldsymbol{l}_i^s, i = 1, 2, \ldots, N^s\}$ and $\mathcal{L}^p = \{\boldsymbol{l}_j^p, j = 1, 2, \ldots, N^p\}$ represent the portion of the training set containing all the LLRs labeled as *speech-active* or *speech-pause*, respectively. Here we use super script ^{s} and ^{p} to denote the labels. In discriminative training, VAD performance is directly associated with a designed objective function, which can be optimized within the training data.

A. Minimal Classification Error Training

Minimum classification error (MCE) training [6] is a well known discriminative training approach, which aims at minimizing the misclassification errors over the entire training set. The MCE loss function and can be defined as

$$C(\mathcal{L}; \boldsymbol{w}) \triangleq \frac{1}{N^p} \sum_{j=1}^{N^p} \mathbf{1}\left(\boldsymbol{w}^T \boldsymbol{l}_j^p - \eta\right) + \frac{1}{N^s} \sum_{i=1}^{N^s} \mathbf{1}\left(\eta - \boldsymbol{w}^T \boldsymbol{l}_i^s\right) \tag{6}$$

where $\mathbf{1}(z)$ is an indicator function where for argument z,

$$\mathbf{1}(z) = \begin{cases} 1, & z > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

and can be approximated by a sigmoid function $D(z)$ which could be differentiated during the optimization as

$$\mathbf{1}(z) \approx D(z) \triangleq 1/(1 + \exp(-\beta z)) \tag{8}$$

where β is the rate of decay of the sigmoid function.

Basically, the minimization of MCE loss function in (6) can improve the VAD performance in terms of reduced amount of two types of errors, (e.g., the miss-hit errors and false-alarm errors). However, its relation with the receiver operating characteristics (ROC) curve is implicit. To this concern, an alternative novel criterion is considered here to directly improve ROC performance.

B. Maximal Area Under the ROC Curve Training

The ROC curves are frequently used to completely describe the VAD performance. A ROC curve is drawn by varying the decision threshold to reflect the relationship between *speech-hit rate* (HR1), defined as the fraction of all actual speech frames that are correctly classified as speech-active frames against the *false-alarm rate* (FAR0), defined as the the fraction of all the actual speech-pause (e.g., noise only) frames that are incorrectly classified as speech frames.

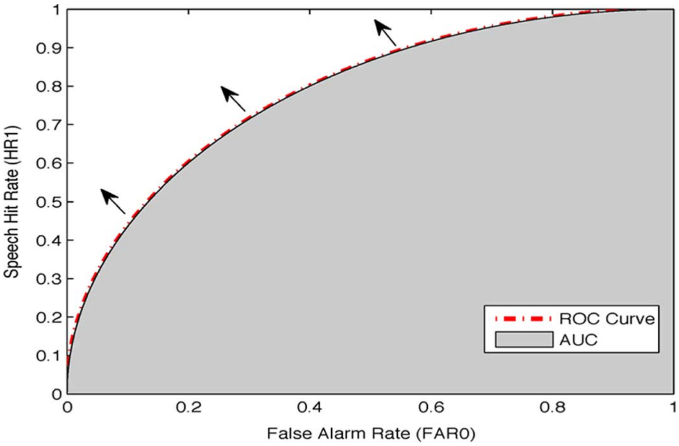

Fig. 1. Illustration of ROC curve and AUC.

Intuitively as illustrated in Fig. 1, the closer the ROC curve is toward the upper left corner, the better the classifier's ability to discriminate between the two classes. Thus, the area under the ROC curve (AUC) is a general, robust measure of classifier discrimination performance, regardless of the decision threshold, which may be unknown, changeable over time, or might vary depending on how the classifier will be used in practical applications.

As shown in [7], the AUC for a binary classifier can be denoted by the value of the normalized Wilcoxon–Mann–Whitney (WMW) statistics at the output of the classifier as

$$A(\mathcal{L}; \boldsymbol{w}) = \frac{1}{N^s \cdot N^p} \sum_{i=1}^{N^s} \sum_{j=1}^{N^p} \mathbf{1}\left(\boldsymbol{w}^T \boldsymbol{l}_i^s - \boldsymbol{w}^T \boldsymbol{l}_j^p\right). \tag{9}$$

Obviously, WMW statistics from a pairwise perspective compares the output of the VAD classifier based on the speech-active LLRs and speech-pause LLRs (e.g., $(\boldsymbol{w}^T \boldsymbol{l}_i^s - \eta) - (\boldsymbol{w}^T \boldsymbol{l}_j^p - \eta) = \boldsymbol{w}^T \boldsymbol{l}_i^s - \boldsymbol{w}^T \boldsymbol{l}_j^p$), and counts on the classification accuracy regardless of the decision threshold η. The larger the WMW statistics, the higher the accuracy of the classification. The optimal weights can be obtained upon the maximization of the WMW statistics. Define a new LLR vector as

$$\Delta \boldsymbol{l}_{i,j} = \boldsymbol{l}_i^s - \boldsymbol{l}_j^p \tag{10}$$

which represents the pairwise difference between the LLRs from the speech-active frame and speech-pause frame. With this, the amount of $\Delta \boldsymbol{l}_{i,j}$ will be $N = N^s \times N^p$. Finally, the MaxAUC loss function can be written as

$$\hat{A}(\mathcal{L}; \boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N^s} \sum_{j=1}^{N^p} D(\boldsymbol{w}^T \Delta \boldsymbol{l}_{i,j}). \tag{11}$$

C. Comparison Between MCE and MaxAUC

As in (6), MCE training requires the decision threshold η to be pre-selected; thereby, optimal weights $\boldsymbol{w}$ could be computed such that the weighted sum of either $\boldsymbol{w}^T \boldsymbol{l}_i^s$ or $\boldsymbol{w}^T \boldsymbol{l}_j^p$

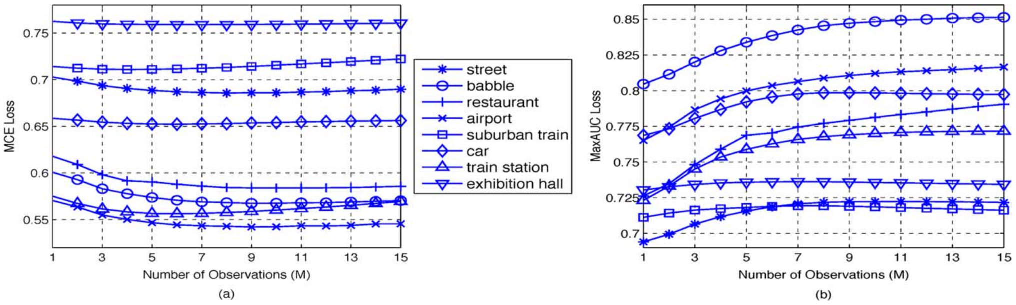

Fig. 2. Optimal values of the loss functions [(a) MCE and (b) MaxAUC] for different number of observations in different AURORA noises.

is separated as far as possible from this particular threshold. However, optimal weights from MaxAUC training are obtained through maximizing the pairwise distance between the speech-active LLRs and speech-pause LLRs, regardless of the decision threshold. Therefore, MaxAUC training has two distinct characteristics: it maximizes the ROC performance on average rather than for a specific threshold; it sufficiently utilizes the training data through a pairwise computation, but in alternative, sacrifices more computation resources, which has to process an amount of $N = N^s \times N^p$ training data compared to the $N^s + N^p$ data for MCE training.

D. Optimization Algorithm

Here, an optimization algorithm is derived for MaxAUC training (noticing that same derivation can be applied to MCE training). As suggested in [4], w_m should satisfy the constraints of $\forall m, w_m \geq 0$ and $\sum_{m=1}^{M} w_m = 1$; hence, a parameter transformation could be employed:

$$\boldsymbol{w} = \tilde{\boldsymbol{w}}^2 \tag{12}$$

where $(\cdot).^2$ denotes an element-wise square operation (i.e., $\forall m, w_m = \tilde{w}_m^2$). This parameter transform automatically guarantees the non-negativeness of $\boldsymbol{w}$. For another constraint of $\sum_{m=1}^{M} w_m = 1$, a corresponding constraint of $\|\tilde{\boldsymbol{w}}\| = 1$ is used, with $\|\cdot\|$ denoting the Euclidean norm (i.e., $\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}}=1$).

The optimization can be formulated in a compact form as

$$\tilde{\boldsymbol{w}}_o = \arg\max_{\|\tilde{\boldsymbol{w}}\|=1} \hat{A}(\mathcal{L}; \tilde{\boldsymbol{w}}). \tag{13}$$

Solving for the optimal weights $\tilde{\boldsymbol{w}}_o$ also leads to optimal $\boldsymbol{w}_o$. Here, (13) is a norm constraint optimization problem with solutions lying in a unit hypersphere. Hence, using an efficient natural gradient algorithm [8], [9], the optimal weights $\boldsymbol{w}$ could be updated with a steepest descent technique as follows:

$$\tilde{\boldsymbol{w}}_{r+1} = \tilde{\boldsymbol{w}}_r + \epsilon\left(I - \tilde{\boldsymbol{w}}_r\tilde{\boldsymbol{w}}_r^T\right)\nabla_{\tilde{\boldsymbol{w}}}\hat{A}|_{\tilde{\boldsymbol{w}}=\tilde{\boldsymbol{w}}_r}, \tag{14}$$

$$\tilde{\boldsymbol{w}}_{r+1} = \tilde{\boldsymbol{w}}_{r+1}/\|\tilde{\boldsymbol{w}}_{r+1}\|, \tag{15}$$

$$\boldsymbol{w}_{r+1} = (\tilde{\boldsymbol{w}}_{r+1}).^2 \tag{16}$$

where $\epsilon > 0$ is the learning rate for updating and r is the iteration index; I is an identity matrix and $\nabla_{\tilde{\boldsymbol{w}}}\hat{A}$ is the gradient of the WMW statistics and could be obtained as

$$\nabla_{\tilde{\boldsymbol{w}}}\hat{A} = \frac{1}{N}\sum_{i=1}^{N^s}\sum_{j=1}^{N^p} 2\beta D(\boldsymbol{w})(1 - D(\boldsymbol{w}))(\tilde{\boldsymbol{w}}. * \Delta\boldsymbol{l}_{i,j}), \tag{17}$$

where $.*$ denotes an element-wise multiplication. The parameter transform step in (12) and (16), and the normalization step in (15) guarantee the constraints on $\boldsymbol{w}$ are satisfied throughout the iterations. Here, the natural gradient is employed due to its optimality for the increasing direction on a hypersphere.

IV. Evaluations

A. Implementation

The proposed VAD is evaluated using an analysis window time-frame of 32 ms, with a 50% overlapping frame for recordings at an 8 kHz sample rate. The STFT is calculated using a Hamming window with an FFT length of 256. The IMCRA [10] and Ephraim–Malah [11] estimators are used for noise power estimation and LLR computation. For all the training, the combination weights are uniformly initialized and $\beta = 1$ for the sigmoid function defined in (8); the learning rate is set as $\epsilon = 0.5(1 - r/1000)$ with maximal iteration of r set to 1000 for MCE and 300 for MaxAUC, respectively.

B. Evaluation for AURORA Data

In this section, the relation between the number of multiple LLRs, M, and the VAD performance is studied using the AURORA database [12]. Here, six different noisy scenarios are considered, with each having nonoverlapping 10 minutes of training data and 10 minutes of test data. The percentage of hand-marked speech-active frames is 53.4%. The signal-to-noise ratio (SNR) is about 5 dB.

Fig. 2(a) shows how the misclassification errors at a pre-selection threshold (e.g., $\eta = 2.1$) drops when M increases. However, excessive observations will not necessarily improve performance at that selected threshold. The optimal M varies for different noise types as well as the decision threshold. The AUC loss shown in Fig. 2(b) gives a more meaningful performance measure regardless of a particular threshold; here, the optimal M that gives the best results could be chosen. As a example, Fig. 3 illustrates the ROC curves for different values of M evaluated for the babble noise.

C. Evaluation for In-Vehicle UTDrive-Noise Data

In this section, the two discriminative training methods are compared. The noises are chosen from the in-vehicle UTDrive-Noise database [13], which consists of diverse noises from 30 different vehicles (including five trucks, five SUVs, and 20 cars) at various driving scenarios. Clean speech is selected from the TIDigits database [14]. Nonoverlapping 10 minutes of training data and 10 minutes of testing data are used for evaluation in each of the vehicle types (e.g., truck, SUV, and car) and the

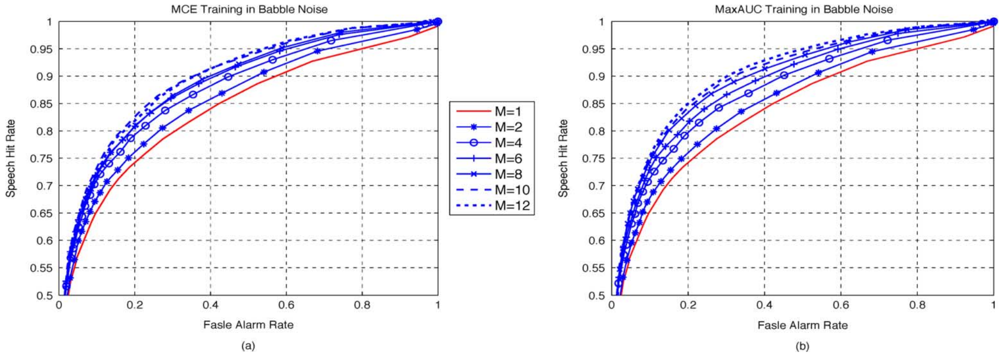

Fig. 3. (a) MCE with $\eta = 2.1$ and (b) MaxAUC ROC curves for different number of observations in the AURORA babble noises.

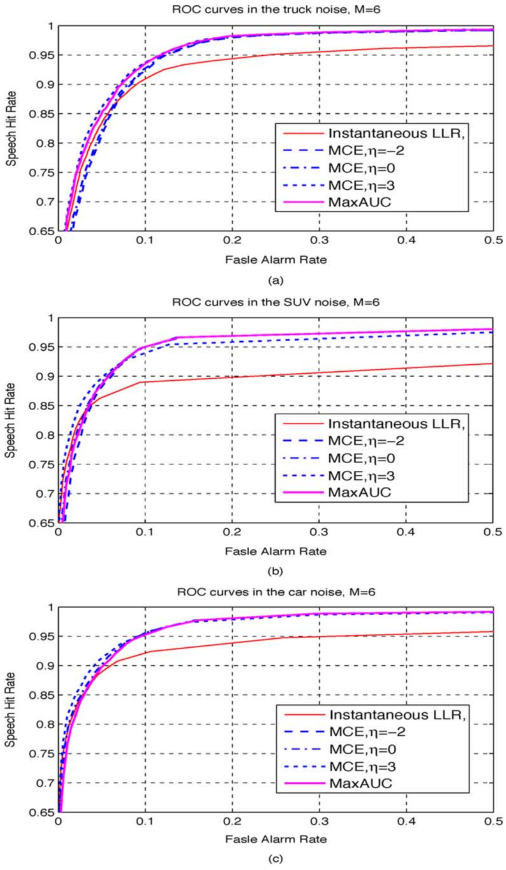

Fig. 4. ROC curves for UTD-In-Vehicle-Noise database: (a) truck, (b) SUV, and (c) car.

percentage of hand-marked speech-active frames is 47.8%. The overall SNR is about 0 dB.

Fig. 4 shows the ROC curves for three different sized vehicles [(a) trucks, (b) SUV and (c) car], when $M = 6$ observation LLRs are used. Both MCE training and MaxAUC training can significantly enhance VAD performances in all vehicle noise types versus the baseline VAD (e.g., $M = 1$, for instantaneous LLR). However, MCE training needs experimental selection of a working threshold through investigation of various ROC curves while MaxAUC training guarantees a global enhancement of ROC curve.

V. Conclusion

In this study, two discriminative training methods for multiple observation based VAD is investigated and evaluated in various noisy environments. A significant VAD improvement was achieved using an automatically obtained environment-dependent weights that can effectively combine the LLRs from multiple observations.

References

[1] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[2] Y. D. Cho, K. Al-Naimi, and A. Kondoz, “Improved voice activity detection based on a smoothed statiscial likelihood ratio,” in *Proc. ICASSP*, 2001, vol. 2, pp. 737–740.

[3] J. Ramírez, J. Segura, C. Beníez, L. García, and A. Rubio, “Statistical voice activity detection using a multiple observation likelihood ratio test,” *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, Oct. 2005.

[4] S.-I. Kang, Q.-H. Jo, and J.-H. Chang, “Discriminative weight training for a statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 15, pp. 170–173, 2008.

[5] J.-H. Chang, N. S. Kim, and S. K. Mitra, “Voice activity detetion based on multiple statistical models,” *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[6] B. H. Juang and S. Katagiri, “Discriminative learning for minimum error classification,” *IEEE Trans. Signal Process.*, vol. 40, no. 12, pp. 3043–3054, Dec. 1992.

[7] L. Yan, R. Dodier, M. C. Mozer, and R. Wolniewicze, “Optimizing classifier performance via an approximation to the Wilcoxon–Mann–Whitney statistic,” in *Proc. ICML*, 2003, pp. 848–855.

[8] S. Amari, “Natural gradient works efficiently in learning,” *Neural Comput.*, vol. 10, no. 2, pp. 251–276, Feb. 1998.

[9] S. C. Douglas, S. Amari, and S. Y. Kung, “On gradient adaptation with unit-norm constraints,” *IEEE Trans. Signal Process.*, vol. 48, no. 6, pp. 1843–1847, Jun. 2000.

[10] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averagings,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[11] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 12, pp. 1109–1121, Dec. 1984.

[12] H. G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condidions,” in *ISCA ITRW ASR2000*, Sep. 2000.

[13] [Online]. Available: http://www.utdallas.edu/research/utdrive/UT-Drive-Website.htm

[14] R. Leonard, “A database for speaker independent digit recognition,” in *Proc. ICASSP*, 1984.