MINIMUM MEAN-SQUARED ERROR A POSTERIORI ESTIMATION OF HIGH VARIANCE VEHICULAR NOISE

Bowon Lee [*]

Hewlett-Packard Laboratories
1501 Page Mill Rd.
Palo Alto, CA 94304
bowon.lee@hp.com

Mark Hasegawa-Johnson

University of Illinois at Urbana-Champaign
Electrical and Computer Engineering
405 N. Mathews Ave,. Urbana, IL 61801
jhasegaw@uiuc.edu

ABSTRACT

In this paper, we describe a method of minimum mean-squared error (MMSE) *a posteriori* estimation of high variance vehicular noise. The proposed method considers spectral instances of noise as sampled values from a stochastic noise process and estimates them with given statistical properties of noise and current noisy observation. Accuracy of the noise estimation method is evaluated in terms of the accuracy of a spectrum-based voice activity detection, in which speech presence is determined by the *a priori* and *a posteriori* signal-to-noise ratios (SNRs) in each frequency bin. VAD experiments are performed on clean speech data by adding four different types of vehicular noise, each with the SNR varying from -10 to 20 dB. Isolated digit recognition experiments are performed using original noisy recordings from the AVICAR corpus. Experimental results show that the proposed noise estimation method outperforms both the MMSE *a priori* noise estimation and autoregressive noise estimation methods especially for low SNR.

1. INTRODUCTION

Speech processing systems such as speech coding and automatic speech recognition are typically designed for clean speech signals as input. In many practical situations, speech is corrupted by background noise. Noisy speech signals are detrimental to speech processing systems, which require speech enhancement algorithms so that speech processing systems perform as they are designed. Speech enhancement algorithms often depend on the existence of a robust voice activity detector (VAD). Even without speech enhancement, a good VAD can substantially improve the word error rate of automatic speech recognition in noise (e.g., [1]). VAD can be modeled as a log-likelihood test, evaluating the relative likelihoods of speech presence vs. absence [2], or as an explicit computation of speech presence probability [3, 4]. Statistical VAD algorithms are mostly based on the signal-to-noise ratio (SNR) [5, 6, 7], thus accurate estimation of the noise power in each frame is critical.

Most systems depend on an MMSE estimate of the noise power spectrum, i.e., an estimate of the expected value of the noise power in each bin of the short time Fourier transform (STFT). The expected noise power may be estimated from the first several frames of a recording, if the first frames are known to contain no speech. Alternatively, the noise power may be recursively updated: Sohn and Sung [6] proposed an autoregressive noise adaptation method with a variable adaptation coefficient that depends on an estimate of the speech presence probability.

With higher noise power, the noise spectrum has higher variance, and therefore, even though the noise is stationary, the MMSE estimate of the noise spectrum may not be close to the noise spectrum of the current observation. High noise power also disrupts autoregressive noise adaptation methods, because these methods depend on an estimate of speech presence probability: in low SNR recordings, estimated speech presence probabilities are inaccurate. For these reasons, speech enhancement and VAD algorithms that perform well with high SNR may nevertheless fail in low SNR environments such as an automobile.

This paper proposes an MMSE *a posteriori* estimation of noise based on the MMSE *a priori* estimation of noise, combined with a current noisy observation, employing speech presence uncertainty. The proposed method treats spectral instances of noise as independent and identically distributed (IID) sampled values of a random variable; thus, unlike noise adaptation methods, the proposed method does not assume that noise spectral amplitude is predictable from its own recent history. Experimental results show that the proposed noise estimation achieves higher VAD accuracy in an automotive environment compared to MMSE *a priori* noise estimation and autoregressive noise adaptation methods.

[*] Bowon Lee performed the work while at the University of Illinois at Urbana-Champaign. This work was funded by Motorola Corporation, and by NSF grant 05-34106.

2. BACKGROUND: VOICE ACTIVITY DETECTION

2.1. Statistical Noise Model

Consider an input signal x consisting only of stationary noise n and assume that noise is a random process with an unknown probability density function (pdf) with zero mean. Let the short-time Fourier transform STFT of x be given by

$$X_k^m = \sum_{n=0}^{L-1} x[n+mL]e^{-j\frac{2\pi kn}{N}} \tag{1}$$

If we consider that the STFT coefficients of x are a weighted sum of samples of the corresponding random process, then according to the central limit theorem, as $L \to \infty$, the STFT coefficients X_k^m asymptotically have Gaussian pdf with zero mean [3]. Thus, the pdf of the k^{th} frequency bin, X_k^m, can be expressed as

$$p(X_k^m) = \frac{1}{\pi\lambda_N(k)} \exp\left\{-\frac{|X_k^m|^2}{\lambda_N(k)}\right\} \tag{2}$$

where $\lambda_N(k) = E[|X_k^m|^2]$ denotes the noise variance.

We can show that the power at the k^{th} spectral component $|X_k^m|^2$ has an exponential pdf with mean $\lambda_N(k)$. Thus, the variance of DFT of noise $\lambda_N(k)$ is equivalent to the MMSE estimation of noise power.

Figure 1 depicts the histogram of squared STFT amplitudes in one frequency bin. The signal being transformed is white Gaussian noise. The dashed line is an exponential pdf; the solid line is a normalized histogram of the squared amplitude of the 100^{th} frequency bin of a 400 point DFT of white Gaussian noise.

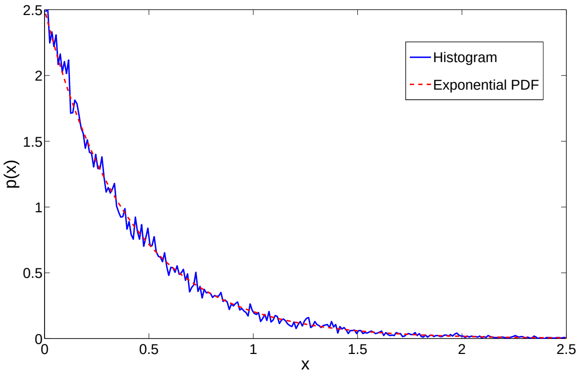

Fig. 1. Histogram of Spectrum versus an Exponential pdf

2.2. Voice Activity Detection

Assume now that the measurement x may contain speech, i.e., it may be the case that $x = s + n$. Voice Activity Detection (VAD) compares the probabilities of two hypotheses:

$$\begin{cases} H_0 : X_k^m = N_k^m, & \text{speech absent} \\ H_1 : X_k^m = S_k^m + N_k^m, & \text{speech present} \end{cases} \tag{3}$$

where S_k^m, N_k^m, and X_k^m are K-dimensional STFT vectors of speech, noise, and noisy speech respectively. The pdf of X_k^m given H_0 is equivalent to Eq. (2) and the pdf given H_1 is [3]

$$p(X_k|H_1) = \frac{1}{\pi(\lambda_N(k)+\lambda_S(k))} \exp\left\{-\frac{|X_k|^2}{\lambda_N(k)+\lambda_S(k)}\right\}$$

where $\lambda_S(k) = E[|S_k^m|^2]$ and $\lambda_N(k) = E[|N_k^m|^2]$ denote the speech and noise variance respectively. The likelihood ratio at the k^{th} frequency bin is

$$\Lambda_k = \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1+\xi_k} \exp\left\{\frac{\gamma_k\xi_k}{1+\xi_k}\right\} \tag{4}$$

where $\xi_k = \lambda_S(k)/\lambda_N(k)$ and $\gamma_k = |X_k^m|^2/\lambda_N(k)$ are defined as *a priori* and *a posteriori* SNR respectively [3].

3. BACKGROUND: NOISE ESTIMATION

3.1. MMSE *a priori* Noise Estimation

In practice, we do not have an infinite length noise sequence. The most common method of noise estimation given a finite length noise sequence is periodogram estimation.

$$\hat{\lambda}_N^m(k) = |X_k^m|^2 \tag{5}$$

where X_k^m is the STFT of noise only signal x in the m^{th} frame as defined in Eq. (1).

$\hat{\lambda}_N^m(k)$ is exponentially distributed, so its expectation is also its standard deviation. We can use Bartlett's procedure to reduce the variance of $\hat{\lambda}_N^m(k)$ by averaging M frames.

$$\bar{\lambda}_N(k) = \frac{1}{M}\sum_{m=0}^{M-1} \hat{\lambda}_N^m(k) \tag{6}$$

This method requires a length LM sequence of noise only observations. $\bar{\lambda}_N(k)$ is an unbiased and consistent estimator of $\lambda_N(k)$:

$$E\left[\bar{\lambda}_N(k)\right] = \lambda_N(k) \tag{7}$$

$$E\left[\left(\bar{\lambda}_N(k) - \lambda_N(k)\right)^2\right] = \frac{1}{M}\lambda_N(k)^2 \tag{8}$$

However, Eqs. (7) and (8) do not imply that $\bar{\lambda}_N(k)$ predicts any particular instance of $|N_k^m|^2$ with high accuracy: $|N_k^m|^2$ is exponentially distributed, so its standard deviation equals its mean.

3.2. MMSE *a posteriori* Noise Estimation

Considering the speech presence uncertainty, the MMSE estimate of the noise at the k^{th} frequency bin in the m^{th} frame given current noisy observation is [2]

$$\begin{aligned}\hat{\lambda}_N^m(k) &= E\big[|N_k^m|^2 \big| X_k^m\big] \\ &= E\big[|N_k^m|^2 \big| H_0\big] p(H_0 \big| X_k^m) \\ &\quad + E\big[|N_k^m|^2 \big| H_1\big] p(H_1 \big| X_k^m)\end{aligned} \tag{9}$$

Using Bayes' rule:

$$\begin{aligned}p(H_0 \big| X_k^m) &= \frac{p(X_k^m \big| H_0)p(H_0)}{p(X_k^m \big| H_0)p(H_0) + p(X_k^m \big| H_1)p(H_1)} \\ &= \frac{1}{1+\epsilon\Lambda_k^m}\end{aligned} \tag{10}$$

where $\epsilon = p(H_1)/p(H_0)$ and $\Lambda_k^m = p(X_k^m|H_1)/p(X_k^m|H_0)$ is the likelihood ratio of the m^{th} frame as in (4). We can derive $p(H_1|X_k^m)$ similarly,

$$p(H_1|X_k^m) = \frac{\epsilon\Lambda_k^m}{1+\epsilon\Lambda_k^m} \tag{11}$$

If we let $\beta_k^m = p(H_1 \big| X_k^m) = \epsilon\Lambda_k^m/(1+\epsilon\Lambda_k^m)$ and substitute (10) and (11) into (9), then

$$\hat{\lambda}_N^m(k) = \beta_k^m E\big[|N_k^m|^2 \big| H_1\big] + (1-\beta_k^m) E\big[|N_k^m|^2 \big| H_0\big] \tag{12}$$

3.3. Autoregressive Noise Adaptation

In (12), we need the estimates of noise spectrum of each hypothesis, $E\big[|N_k^m|^2 \big| H_0\big]$ and $E\big[|N_k^m|^2 \big| H_1\big]$. Sohn and Sung [6] proposed that, under hypothesis H_0, we can use the current noisy observation, i.e.,

$$E\big[|N_k^m|^2 \big| H_0\big] = |X_k^m|^2 \tag{13}$$

Under hypothesis H_1, $|X_k^m|^2$ contains speech as well as noise, and is therefore not an accurate estimate of the noise power. Assuming that the VAD probability β_k^m has been correctly estimated in all previous frames, the best available estimate of the noise power is therefore

$$E\big[|N_k^m|^2 \big| H_1\big] = \hat{\lambda}_N^{m-1}(k) \tag{14}$$

Combining Eqs. (12) through (14) yields

$$\hat{\lambda}_N^m(k) = \beta_k^m \hat{\lambda}_N^{m-1}(k) + (1-\beta_k^m)|X_k^m|^2 \tag{15}$$

Sohn and Sung [6] proposed that, if β_k^m is an accurate estimate of the speech presence probability in each frame, then Eq. (15) is an equally accurate estimate of the noise power in the m^{th} frame. Under these circumstances, $\hat{\lambda}_N^m(k)$ takes into account all information about the underlying noise process that can be extracted from frames up to and including the current frame.

4. PROPOSED NOISE ESTIMATION METHOD

The autoregressive noise estimator $\hat{\lambda}_N^m(k)$ proposed in Eq. (15) is optimal, if and only if the speech presence probability estimate β_k^m is accurate. Unfortunately, in a low SNR environment, β_k^m is itself a random variable with high variance. β_k^m is a sigmoid transformation of the random variable $|X_k^m|^2$:

$$\beta_k^m = \frac{e^{|X_k^m|^2/(a_k\lambda_N(k))}}{(a_k/\epsilon) + e^{|X_k^m|^2/(a_k\lambda_N(k))}} \tag{16}$$

where $a_k = (1+\xi_k)/\xi_k$.

The input threshold of the sigmoid—the value of $|X_k^m|^2$ at which $\beta_k^m = 0.5$—is given by $\theta_k = a_k\lambda_N(k)\log\left(\frac{a_k}{\epsilon}\right)$. Any noise-only frame in which $|N_k^m|^2 > \theta_k$ will cause a "false positive:" $\beta_k^m \approx 1$ despite the absence of speech. Eq. (15) prohibits these false positives from contributing to the autoregressive estimate $\hat{\lambda}_N^m(k)$, therefore, over time, the estimate $\hat{\lambda}_N^m(k)$ tends to under-estimate the true expected value $E[|N_k^m|^2]$, and to over-estimate the probability of speech presence in any given frame.

In order to more precisely estimate the amount by which autoregressive noise adaptation under-estimates $\lambda_N(k)$ in low-SNR environments, let us treat β_k^m as a binary random variable—a unit step function of $|N_k^m|^2$, rather than a sigmoid function. Define $\rho = P\left(\beta_k^m \geq 0.5\right)$; under the assumption, $E\left[\beta_k^m\right] = \rho$. By integrating the pdf of $|N_k^m|^2$, we find that

$$\rho = \int_{a_k\log(a_k/\epsilon)}^{\infty} e^{-t}dt = \left(\frac{a_k}{\epsilon}\right)^{-a_k} \tag{17}$$

In terms of ρ, the expected value of $\hat{\lambda}_N^m(k)$ is approximately

$$E\left[\hat{\lambda}_N^m(k)\right] \approx \rho E\left[\hat{\lambda}_N^{m-1}(k)\right] + (1-\rho)\, E\left[|X_k^m|^2 \big| \beta_k^m < 0.5\right]$$

However,

$$\begin{aligned}&(1-\rho)\, E\left[|X_k^m|^2 \big| \beta_k^m < 0.5\right] \\ &= \lambda_N(k)\int_0^{a_k\log(a_k/\epsilon)} te^{-t}dt \\ &= \lambda_N(k)\left[1-\rho+\rho\log\rho\right]\end{aligned} \tag{18}$$

Combining equations, we find that

$$E\left[\hat{\lambda}_N^m(k)\right] \approx \rho\hat{\lambda}_N^{m-1}(k) + \lambda_N(k)\left[1-\rho+\rho\log\rho\right] \tag{19}$$

If we begin with a perfect estimate $\hat{\lambda}_N^1(k) = \lambda_N(k)$, then Eq. (19) demonstrates that $\hat{\lambda}_N^m(k)$ will tend to decay over time, with an initial slope of $\rho\log\rho\lambda_N(k) < 0$. The scaling factor $\rho\log\rho$ is most negative at values of $\rho = e^{-1}$; for example, if $\epsilon = 1$, then the smallest value of $\rho\log\rho$ (and therefore, according to the estimate in Eq. (19), the worst under-estimation of $\lambda_N(k)$ by the autoregressive estimator) occurs at an SNR of $\xi_k = 1.3$, quite close to 0 dB SNR.

Table 1. Four noise conditions from the AVICAR database

Condition	Description
35U	Car running at 35 mph with windows closed
35D	Car running at 35 mph with windows open
55U	Car running at 55 mph with windows closed
55D	Car running at 55 mph with windows open

In high-noise environments, therefore, error propagation using Eq. (15) is an important problem. Error propagation can be avoided by applying a certain amount of prior knowledge to the problem. For example, if the noise process is known to be stationary, and if the first M frames of the signal are known to contain no speech, then an *a priori* periodogram estimate $\bar{\lambda}_N(k)$ of $E[|N_k^m|^2]$ with known standard error may be computed using Eq. (6). If we assume that intervening frames provide no further information about $E[|N_k^m|^2]$, then

$$E\left[|N_k^m|^2 \middle| H_1\right] = \bar{\lambda}_N(k) \tag{20}$$

and Eq. (12) becomes

$$\hat{\lambda}_N^m(k) = \beta_k^m \bar{\lambda}_N(k) + (1 - \beta_k^m)|X_k^m|^2 \tag{21}$$

If it is highly likely that the speech is present, i.e., $\Lambda_k^m \gg 1$, thus $\beta_k^m \approx 1$, then Eq. (21) sets $\hat{\lambda}_N^m(k)$ equal to the mean estimate of the noise spectrum. Thus the method proposed in Eq. (21) is subject to false-positive errors, just like the autoregressive estimator, but Eq. (21) does not propagate error. Instead, a false-positive frame is treated just like any other frame about which we have no certain knowledge of the noise spectrum: the noise estimate is backed off to the *a priori* noise estimator $\bar{\lambda}_N(k)$.

The proposed noise spectrum estimation method can be interpreted as an *a posteriori* MMSE estimate of the noise power in the current frame, when the noise process is stationary but with high variance. Experimental results show that the proposed noise estimation method provides higher accuracy especially for low SNR cases.

5. EXPERIMENTS

Evaluation included two experimental tests. First, noise and speech were electronically mixed, and the noise estimation methods in Eqs. (6), (15), and (21) were tested in the task of voice activity detection (VAD). Second, original noisy speech data were endpointed using each of the three VADs, and word error rates (WER) were computed using mixture Gaussian HMMs.

VAD tests used 62 sentences from the TIMIT database [8]. One second of silence was inserted between adjacent sentences, making a total duration of 212 seconds, of which 54% seconds contain speech. Frame duration chosen for experiments is 10 ms and each frame is marked as either speech or silence according to the transcription by marking frames with speech in more than 50% of their duration to be speech, and silence otherwise. We added four different car noises out of five from the AVICAR database [9] by increasing SNR's from -10 dB to 20 dB with 5 dB increments for each noise condition. AVICAR (http://www.ifp.uiuc.edu/speech/AVICAR/) is a database of multi-camera, multi-microphone audiovisual speech acquired from 100 talkers in moving cars. The best audio-only isolated digit word error rate (WER) achieved on this corpus, averaged across all noise conditions, is 3.95%, using a missing-features approach called the phoneme restoration HMM [1]. The best video-only isolated digit WER on this corpus is 62.5% [10]. Audiovisual error rates have not improved upon audio-only rates for this corpus.

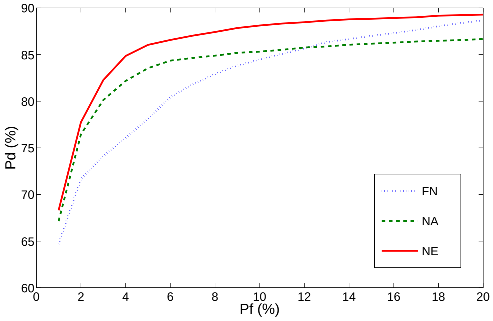

Fig. 2. ROC curve for 5 dB SNR with 55U noise condition, FN: Fixed noise, NA: Autoregressive Noise adaptation, NE: Proposed noise estimation

Description of the noise conditions extracted from the database are listed in Table 1. We used the initial 20 frames (200 ms) for the mean estimate of noise assuming that they contain only noise.

Performance of VAD with different noise estimation methods are compared by correct speech detection and false alarm probabilities (P_d and P_f). Figure 2 depicts the receiver operation characteristics (ROC) for 5 dB SNR with noise condition 55U for three different noise estimation methods. "Fixed noise" estimation (FN) refers to Eq. (6); "Noise adaptation" (NA) refers to Eq. (15); "Noise estimation" (NE) refers to the proposed backoff method, Eq. (21).

In order to quantitatively present the accuracies of different noise estimation methods, we set the threshold η such that $P_f = 5\%$ and compared P_d for each noise estimation method. Summary of results across all noise conditions according to the SNR's is in Table 2. We can see from Table 2 that the soft-decision based noise adaptation method performs worse than the fixed noise spectrum when the SNR is lower than 0 dB, which illustrates that with low SNR, the noise estimate in the previous frames are significantly different from the current noise spectrum, thus autoregressive adaptation makes VAD

Table 2. Summary of P_d's of the VAD's ($P_f = 5\%$). FN="Fixed Noise" (Eq. 6), NA="Noise Adaptation" (Eq. 15), NE=proposed "Noise Estimation" method (Eq. 21).

	NE	NA	FN
SNR	P_d (%)	P_d (%)	P_d (%)
-10 dB	35.89	25.76	28.64
-5 dB	55.41	45.65	47.31
0 dB	72.73	65.69	64.76
5 dB	84.11	81.12	77.88
10 dB	90.74	89.83	86.99
15 dB	93.47	92.90	91.91
20 dB	94.28	94.14	94.09
Overall	75.23	70.73	70.22

Table 3. WER in percent, HMM isolated digit recognizers, noisy speech AVICAR recordings, five-fold cross-validation. BF=beamformed audio, no VAD; NE=BF+noise estimation VAD, NA=BF+noise adaptation VAD, FN=BF+fixed-noise VAD

Noise Condition	BF	NE	NA	FN
IDL	3.41	2.84	2.84	3.13
35U	2.30	2.80	2.86	3.47
35D	2.40	3.64	3.75	4.34
55U	3.51	4.32	4.79	5.38
55D	6.02	7.00	8.41	9.21
Overall	3.49	4.07	4.47	5.04

perform worse. The VAD with the proposed noise estimation method has higher accuracy compared to the other two noise estimation methods especially for low SNR.

The three VADs described in Fig. 2 and Table 2 were used to endpoint isolated digit utterances prior to automatic speech recognition (ASR). In this experiment, all audio signals are original recordings from the AVICAR corpus, therefore "ground truth" for the VAD is unknown (there is no ground truth specification of the beginning or ending of speech in each file). ASR was conducted using mixture Gaussian hidden Markov models (HMMs). HMMs were trained on data from 60 talkers (all noise conditions), the number of Gaussians was increased until error rate reached a minimum on data from 20 talkers, and then error rate was computed for the remaining 20 talkers. Five-fold cross validation was used to compute the word error rates (WER) in Table 3. WERs reported for each noise condition each summarize 2000 tokens, with a maximum WER of about 5%, therefore WER differences of 0.8% are significant at the $p = 0.05$ level; WERs reported in the final row of each table summarize 10000 tokens, therefore WER differences of 0.35% are significant. All audio signals are the result of delay-and-sum beamforming using a 7-microphone horizontal array.

6. CONCLUSION

In this paper, we proposed a MMSE *a posteriori* noise estimation method by considering the instantaneous noise spectrum as sampled values from an underlying noise process. Experimental results show that the proposed noise estimation method provides higher accuracy for VAD and for isolated digit recognition than the fixed mean estimate or noise adaptation methods with high variance vehicular noise.

7. REFERENCES

[1] B. Lee, *Robust Speech Recognition in a Car Using a Microphone Array*, Ph.D. thesis, University of Illinois at Urbana-Champaign, 2006.

[2] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, pp. 137–145, 1980.

[3] Y. Ephraim and D. Malah, "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, 1984.

[4] I.-Y. Soon, S.-N. Koh, and C.-K. Yeo, "Improved noise suppression filter using self-adaptive estimator of probability of speech absence," *Sig. Process.*, vol. 75, no. 2, pp. 151–159, 1999.

[5] J. Sohn, N.-S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Sig. Process. Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[6] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," *Proc. Int. Conf. Acoust., Speech, and Sig. Process.*, pp. 365–368, 1998.

[7] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Sig. Process. Letters*, vol. 8, no. 10, pp. 276–278, 2001.

[8] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Comm.*, vol. 9, no. 4, pp. 351–356, 1990.

[9] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: Audio-visual speech corpus in a car environment," *Proc. Int. Conf. Spoken Lang. Process.*, 2004.

[10] Y. Fu, X. Zhou, M. Liu, M. Hasegawa-Johnson, and Thomas Huang, "Lipreading by locality discriminant graph," in *Proc. ICIP*, 2007.