A PHONEMIC RESTORATION APPROACH FOR AUTOMATIC SPEECH RECOGNITION WITH HIGHLY NONSTATIONARY BACKGROUND 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
405 N. Mathews Ave.
Urbana, IL 61801
jhasegaw@uiuc.edu

ABSTRACT

Automatic speech recognition for human computer interaction in vehicular environments is a challenging task due to existence of highly nonstationary background noise. At moderate signal-to-noise ratio ranging between 0 and 20 dB, word error rate can be dramatically reduced using methods such as speech enhancement and robust feature extraction. When a burst of nonstationary noise overlaps speech duration, however, the segmental signal-to-noise ratio often drops well below 0 dB making conventional methods almost ineffective. For this scenario, we propose a method exploiting the phonemic restoration effect to reduce word error rate. In this method we use a microphone array to determine speech duration corrupted by presence of nonstationary noise and mark features from that duration as unreliable. Then we selectively discard them, or reconstruct underlying speech features by examining adjacent features marked as reliable. Experimental results on the AVICAR database show that our proposed method reduce word error rate in a statistically meaningful sense.

1. INTRODUCTION

Performance of automatic speech recognition (ASR) is highly dependent upon type mismatch between test and training environments because speech models are mostly trained with clean speech data [1]. For ASR systems in noisy environments, the most challenging task is to match the trained speech models by extracting reliable speech features from noisy speech observations. For this, speech enhancement methods such as spectral subtraction [2] and minimum mean-square error (MMSE) spectral amplitude estimation [3] methods have been widely used in the literature. Recently, the missing feature approach [4], which determines reliable speech components from noisy speech observation in the sectary-temporal domain, has been reported to provide better performance than speech enhancement methods [5]. This method, however, needs to determine speech features from noisy observation using computational auditory scene analysis (CASA) [6] and thus requires huge computational cost and has limitation to use only the speech spectrum as features for ASR.

Use of a microphone array makes it possible to do spatial filtering of the input signals by employing various adaptive beamforming techniques [7, 8]. Unfortunately, factors such as correlated noise, steering delay errors, inhomogeneous microphones characteristics, and echoes of source speech significantly degrade the performance of the adaptive beamforming algorithms [9]. For these reasons, it has been reported that using a nonadaptive delay-and-sum beamformer performs better than adaptive beamformers in terms of word error rate (WER) of ASR [10].

Existing methods for robust ASR still provide poor performance especially when the signal-to-noise (SNR) is low (below 0 dB) and/or background noise is highly nonstationary, which are common cases for in-vehicle applications. With highly nonstationary noise, the extracted speech features may still contain significant noise components either in noise-only segments or speech segments. The former appear as nonexistent speech features and the latter as distorted speech features to ASR systems, both of which are detrimental to the ASR performance. We can employ voice activity detection (VAD) to use speech features only. At low SNR, VAD still can be accomplished by using statistical models of the speech and noise [11]. This method, however, is not effective when nonstationary noise power is comparable to the speech power i.e., segmental SNR is at or below 0 dB.

In order to address this scenario, we propose a phonemic restoration approach for robust ASR. The phonemic restoration is a phenomenon in which humans claim to hear missing phonemes even though the phonemes have been replaced with noise having the same power as the replaced speech [12]. This can be considered as an auditory counterpart of the so-called “picket fence” effect of human tendency to visually reconstruct a scene obstructed by a picket fence [13]. The proposed phonemic restoration hidden Markov model (PRHMM) operates on a three-hypothesis VAD scheme consisting of H_0: noise only region, H_S: speech dominant speech region, and H_N: noise dominant speech region. Then we use features from hypotheses H_S and H_N for ASR. For features from H_N, we selectively discard, or reconstruct underlying speech features by examining neighboring features belong to H_S.

This paper is organized as follows. Section 2 describes a three-hypothesis VAD scheme with a microphone array. Section 3 describes the proposed PRHMM method. Section 4 presents experimental results followed by conclusion in Section 5.

2. THREE-HYPOTHESIS VOICE ACTIVITY DETECTION

The proposed three-hypothesis VAD for the PRHMM is structured as a cascade of two VAD methods. First is a spectrum-based VAD (SVAD) with the Gaussian assumption of speech and noise spectra [11], which is followed by a location based VAD (LVAD) exploiting prior knowledge of source location information.

[*] 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.

2.1. Spectrum-Based Voice Activity Detection

For the SVAD, we formulate two hypotheses: H_0 denoting stationary noise only and H_1 denoting stationary noise plus nonstationary signals, consisting of nonstationary noise and/or speech. Then we choose a hypothesis with higher posterior probability given observation $\mathbf{X}$:

$$p(H_1|\mathbf{X}) \underset{H_0}{\overset{H_1}{\gtrless}} p(H_0|\mathbf{X}), \tag{1}$$

where $\mathbf{X} = [X_0, \cdots, X_k, \cdots, X_{K-1}]^T$ is a vector of the K-point discrete Fourier transform (DFT) of the input signal x.

By assuming that the noise is uncorrelated with the speech and the DFT coefficients of speech and noise processes are asymptotically independent Gaussian random variables, we can find the likelihood ratio as [11]

$$\mathcal{L}_k \triangleq \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1+\xi_k} \exp \frac{\gamma_k \xi_k}{1+\xi_k} \underset{H_0}{\overset{H_1}{\gtrless}} \eta_S, \tag{2}$$

where $\xi_k \triangleq \lambda_S(k)/\lambda_N(k)$ and $\gamma_k \triangleq |X_k|^2/\lambda_N(k)$ are *a priori* SNR and *a posteriori* SNR respectively with $\lambda_N(k)$ and $\lambda_S(k)$ denoting noise and speech variances, all at the k^{th} frequency bin [3]. The threshold $\eta_S = p(H_0)/p(H_1)$ can be determined with known prior probabilities or chosen experimentally.

With the SVAD, we assume that the noise variance $\lambda_N(k)$ is slowly varying (quasi-stationary), which is estimated recursively using soft-decision based on speech presence likelihoods [14]. According to Eq. (2) we can see that nonstationary noise and speech both increase γ_k resulting in a bias toward H_1.

2.2. Location-Based Voice Activity Detection

Location-based VAD operates under H_1 determined by the SVAD. Under H_1, it first finds a source location having the maximum steered response power (SRP) and then compute the likelihood of two hypotheses: H_S for speech and H_N non-speech based on prior probability of speech and noise source locations.

For an array of M microphones, the signal at the m^{th} microphone can be expressed as

$$x_m[n] = s[n - \Delta_\mathbf{q}^m] + n_m[n], \quad \text{for } m = 1, 2, \cdots, M \tag{3}$$

where $\Delta_\mathbf{q}^m$ is propagation delay (in samples) determined by the source location $\mathbf{q}$ and a known geometrical microphone array configuration. The output of a filter and sum beamformer steered toward $\mathbf{q}$ can be represented in the DFT domain as

$$Y_\mathbf{q}[k] = \sum_{m=1}^{M} H_m[k] X_m[k] e^{j\omega\Delta_\mathbf{q}^m} \tag{4}$$

where $H_m[k]$ is a filter for the m^{th} microphone. If we choose phase transform (PHAT) filter, i.e., $H_m[k] = 1/X_m[k]$, the output power of the filter-and-sum beamformer is expressed as

$$\mathcal{P}(\mathbf{q}) = \sum_{m=1}^{M} \sum_{l=1}^{M} \sum_{k=0}^{K-1} \frac{X_m[k]X_l^*[k]}{|X_m[k]X_l^*[k]|} e^{j\frac{2\pi k}{K}(\Delta_\mathbf{q}^m - \Delta_\mathbf{q}^l)}. \tag{5}$$

Then we can find the source location which maximizes the SRP as [15]

$$\hat{\mathbf{q}} = \underset{\mathbf{q}\in\mathcal{Q}}{\arg\max}\, \mathcal{P}(\mathbf{q}), \tag{6}$$

where $\mathcal{Q}$ is a search space containing all possible candidate source locations.

If we assume that source and noise location probabilities, $p(\mathbf{q}|H_S)$ and $p(\mathbf{q}|H_N)$ are known *a priori*, we can formulate a likelihood ratio test (LRT) based on $\mathbf{q}$ as

$$\frac{p(\hat{\mathbf{q}}|H_S)}{p(\hat{\mathbf{q}}|H_N)} \underset{H_N}{\overset{H_S}{\gtrless}} \eta_L, \tag{7}$$

where $\eta_L = p(H_N)/p(H_S)$ is a ratio between two prior probabilities, which can be chosen experimentally.

3. PHONEMIC RESTORATION HIDDEN MARKOV MODEL

Phonemic restoration is an effect that humans can still hear the missing phonemes even though they are replaced with noise [12]. One of the mechanisms of the phonemic restoration effect can be described as a "top-down" expectation of the acoustic speech input [16]. In particular, expected language structure such as lexical, semantic, and syntactic information provides some kind of expectation to fill in the missing phonemes. It has also been reported that that the recovered acoustic features of missing phonemes are dependent on the acoustic characteristic of replacing noise [17], which illustrates the "bottom-up" confirmation process in the phonemic restoration effect. In this section, we describe our method to implement both "top-down" and "bottom-up" aspects of the phonemic restoration effect in the HMM framework.

3.1. Observation Probability of PRHMM

For an HMM with N hidden states $S_1, \cdots S_N$, we can train a set of parameters $\lambda = (A, B, \pi)$ where $A = \{a_{ij}\}$ for $a_{ij} = p(q_{t+1} = S_j | q_t = S_i)$ denotes the state-transition probability matrix, $B = \{b_j(O_t)\}$ for $b_j(O_t) = p(O_t|q_t) = S_j$ denotes observation probabilities, $\pi = \{\pi_{S_i}\}$ is a set of prior probability for each state, and q_t denotes a state at time $t \in \{1, 2, \cdots, T\}$. The probability of a sequence of observations $\mathcal{O} = \{\mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_T\}$ can be computed as [18]

$$\begin{aligned} p(\mathcal{O}|\lambda) &= \sum_{\forall Q} p(\mathcal{O}|Q, \lambda) P(Q|\lambda) \\ &= \sum_{\forall Q} \pi_{q_1} b_{q_1}(\mathbf{O}_1) a_{q_1 q_2} b_{q_2}(\mathbf{O}_2) \cdots a_{q_{T-1} q_T} b_{q_T}(\mathbf{O}_T). \end{aligned} \tag{8}$$

The three-hypothesis VAD described in section 2 determined frames belong to either $\mathcal{T}_{H_S}$ or $\mathcal{T}_{H_N}$ corresponding to hypotheses H_S and H_N respectively:

$$\mathcal{T}_{H_S} = \{t_{s_1}, t_{s_2} \cdots t_{s_S}\} \tag{9a}$$

$$\mathcal{T}_{H_N} = \{t_{n_1}, t_{n_2} \cdots t_{n_N}\} \tag{9b}$$

For observation at time $\tilde{t}$ which belongs to the set $\mathcal{T}_{H_N}$, its observation probability $b_{q_{\tilde{t}}}(O_{\tilde{t}})$ does not well represent the observation of the hidden speech state. In this case, we use a compensated observation probability, $\tilde{b}_{q_{\tilde{t}}}(O_{\tilde{t}})$ instead of $b_{q_{\tilde{t}}}(O_{\tilde{t}})$. Then we can rewrite Eq. (8) as

$$P(\mathcal{O}|\lambda) = \sum_{\forall Q} \pi_{q_1} A_Q \prod_{t\in\mathcal{T}_{H_S}} b_{q_t}(\mathbf{O}_t) \prod_{\tilde{t}\in\mathcal{T}_{H_N}} \tilde{b}_{q_{\tilde{t}}}(\mathbf{O}_{\tilde{t}}), \tag{10}$$

where $A_Q \triangleq a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}$.

3.2. Missing Speech Features

This section will describe how we treat missing speech features from noise dominant frames to be useful for ASR. First of all, we need to determine whether or not it is possible to reconstruct the missing speech features. Considering the “bottom-up” confirmation process in phonemic restoration [17], we may be able to reconstruct noise masked speech features as far as the masking noise has similar acoustic features. Unfortunately, this is not guaranteed because we have limited information about the missing phoneme once it is begin masked. Nevertheless, we propose that the features from the missing phoneme are reconstructible if the masking noise has a short duration during which state transition does not occur. To confirm this, we consider the following two constraints to determine the missing part of speech is reconstructible.

1. The total duration of the masked speech frames should be less than a typical duration of individual HMM state.
2. Adjacent unmasked observations should be similar in some sense, i.e. these observations are from the same hidden state.

The first constraint should vary according to specific HMM state and can be chosen experimentally. For the second constraint, we can measure acoustic similarity as distance between feature vectors. In particular, we use a generalized Itakura-Saito distance (GISD), which has been reported to provide higher correlation with human perception with either the MFCC or PLP features than the linear spectral features [19]:

$$\text{GISD}(\mathbf{O}_{t_i}, \mathbf{O}_{t_j}) = \sum_{k=0}^{K-1}\{\frac{O_{t_j}(k)}{O_{t_i}(k)} - \log\frac{O_{t_j}(k)}{O_{t_i}(k)} - 1\} \underset{similar}{\overset{dissimilar}{\gtrless}} \eta, \tag{11}$$

where $O_{t_i}(k)$ and $O_{t_j}(k)$ denote the magnitude of the k^{th} spectral or cepstral coefficients at time t_i and t_j, both belong to $\mathcal{T}_{H_S}$, and K is the number of coefficients for computing the GISD.

If the similarity measure between adjacent unmasked speech features is below η, then we decide that the adjacent observations are from the same HMM state as well as the missing phoneme. Then we reconstruct the missing part of speech by linearly interpolating the adjacent unmasked observations. When the similarity measure between them is above η, then we mark the masked part of speech as unreliable providing no information about the acoustic observation of the missing phonemes. Therefore, we replace observation probability of a missing speech frame in Eq. (10) by marginalizing it over the entire observation space, which is equivalent to 1.

4. EXPERIMENTS

This section presents experiments to evaluate the proposed algorithm. The AVICAR database [20], a multi-microphone database consisting of five different noise conditions (See Table 1.) has been used as training and testing data for ASR experiments.

Table 1. Five noise conditions in the AVICAR database.

Condition	Description
IDL	Car in idling
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

Speech models are trained using the data from the IDL noise condition only because it corresponds to the trained models of “clean speech,” which most of the ASR systems use. Each digit is modeled as a sequence of three HMMs, *silence-digit-silence* by treating individual recordings at each microphone separately. All ASR experiments use a vocabulary with size 11 ($\{$*zero, one, two, ..., nine, oh*$\}$) and each word is modeled as a nine-state HMM with seven emitting states.

Training of the HMMs are done by using the hidden Markov model toolkit (HTK) [21] using the PLP speech features [22] with energy and dynamic features consisting total of 39 acoustic features. Frame size of 25 ms has been chosen for the experiments. All recognition tests are done using Matlab with the models trained by HTK. Recognition tests are done on utterances in all noise conditions using a fivefold cross-validation paradigm, in which the HMM is trained using speech of talkers from the 8 groups out of 10, among which recordings from the 6 groups are used to train the model, and 2 groups are used to evaluate the trained model to choose the best trained HMM. Testing was done using recordings of the remaining talkers from the remaining 2 groups which are not used for training.

ASR tests consist of four experiments. The baseline (BL) is an ASR on the entire duration of the given speech data using a *silence-speech-silence* model without beamforming. The second set of experiments (BF) is an ASR with beamforming with the rest of the settings are same as the baseline. The third set of experiment (BF_VAD) is an ASR with the VAD, using only the detected speech region for ASR using HMM without the silence model. The fourth set of experiment (BF_PR) is an ASR with the proposed PRHMM method. These four experiments are summarized in Table 2

Table 2. List of experiments.

ID	Description
BL	Baseline
BF	Beamforming
BF_VAD	Beamforming and VAD
BF_PR	Beamforming and PRHMM

The PRHMM requires the information of the speech-dominant speech region (H_S) and that of noise-dominant speech region (H_N), which is proposed by the three-hypothesis VAD described in Section 2. The threshold η in Eq. (11) to determine whether the features from $\mathcal{T}_{H_N}$ is reconstructible, is experimentally set to 0.3 and the maximum allowed number of missing frames between neighboring unmasked frame is set to 5. Missing frames satisfying the above constraints are reconstructed by linear interpolation of adjacent unmasked frames belong to $\mathcal{T}_{H_S}$. Experimental results is summarized in terms of the WER in Table 3

Table 3. Summary of the word error rate (%)

Condition	BL	BF	BF_VAD	BF_PR
IDL	4.22	2.28	2.39	2.39
35U	13.16	11.25	7.67	7.67
35D	24.23	13.26	10.15	9.56
55U	21.40	17.48	10.46	10.52
55D	34.95	18.97	15.41	15.10
Overall	19.26	12.52	9.10	8.94

In overall, the decrease of more than 50 % in WER is from

BL to BF_PR has been achieved. According to Table 3 the overall performance increase from VAD-based recognition (BF_VAD) to the PRHMM (BF_PR) is not significant in terms of the WER. According to the McNemar's tests [23], probability P that two outcomes are from the same algorithm between BF_PR and BF_VAD is $P(13, 20) = 0.2962$, which is above the significance level set to $\alpha = 0.02$. Since the phonemic restoration approach is based on the assumption that the nonstationary noises overlaps speech and noise conditions with windows open (35D and 55D) would better correspond to this assumption than other noise conditions. McNemar's tests only for 35D and 55D conditions gives $P(2, 12) = 0.0129$, which is below the significance level.

As illustrated in Table 3, beamforming significantly reduces the WER, especially for 35D and 55D noise conditions, which are supposed to contain noise mainly from locations other than source speech. It is interesting to note that the WER for 35D is higher than 55U in baseline, but after beamforming, WER for 35D becomes lower than that of 55U. This illustrates that beamforming reduces the effect of nonstationary noise which is further improved by the PRHMM. For noise conditions 35U and 55U, WER does not decrease as significantly as 35D and 55D with beamforming, but a decrease of WER becomes significant with VAD. This shows that with existence of nonstationary noise, the VAD tend to have more false alarms of speech presence such that speech recognizer uses nonspeech regions.

5. CONCLUSION

This paper proposed a method for a microphone array-based ASR by integrating the spatial selectivity of a microphone array and the phonemic restoration effect into the HMM framework. ASR experiments on isolated digits showed statistically significant improvement of the proposed method. According to [16], longer words promote stronger phonemic restoration because they have more evidence for restoration. So it would be meaningful to try words having more phonemes than digits. In addition, the proposed method can be extended to the task of continuous speech recognition, where syntactic and semantic information as well as lexical information can contribute the phonemic restoration.

6. REFERENCES

[1] Yifan Gong, "Speech recognition in noisy environments: A survey," *Speech Comm.*, vol. 16, no. 3, pp. 261–291, April 1995.

[2] Steven F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 27, no. 2, pp. 113–120, April 1979.

[3] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[4] Martin Cooke, Phil Green, Ljubomir Josifovski, and Ascension Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, pp. 267–285, 2001.

[5] Bhiksha Raj and Richard M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Process. Magazine*, vol. 22, no. 5, pp. 101–116, September 2005.

[6] Guy J. Brown and Martin Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[7] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propag.*, vol. 30, no. 1, pp. 27–34, 1982.

[8] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *Signal Processing, IEEE Transactions on*, vol. 47, no. 10, pp. 2677–2684, Oct 1999.

[9] M. S. Brandstein and D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, Berlin, Germany, 2001.

[10] Michael L. Seltzer, *Microphone array processing for robust speech recognition*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, July 2003.

[11] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[12] Richard M. Warren, "Perceptual restoration of missing speech sounds," *Science*, vol. 167, pp. 392–393, 1970.

[13] G. A. Miller and J. C. R. Licklider, "The intelligibility of interrupted speech," *J. Acoust. Soc. Am.*, vol. 22, pp. 167–173, 1950.

[14] Bowon Lee and Mark Hasegawa-Johnson, "Minimum mean squared error a posteriori estimation of high variance vehicular noise," in *Biennial on DSP for In-Vehicle and Mobile Systems*, June 2007.

[15] J. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments*, Ph.D. thesis, Brown University, Providence, RI, May 2000.

[16] A. G. Samuel, "Phonemic restoration: Insights from a new methodology," *Journal of Experimental Psychology: General*, vol. 110, pp. 474–494, 1981.

[17] A. G. Samuel, "The role of bottom-up confirmation in the phonemic restoration illusion," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 7, pp. 1124–1131, 1981.

[18] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[19] J. Wouters and M. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," in *Proc. Int. Conf. Spoken Lang. Process.*, November 1998, vol. 6, pp. 2747–2750.

[20] Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu, and Thomas Huang, "AVICAR: Audio-visual speech corpus in a car environment," in *Proc. Int. Conf. Spoken Lang. Process.*, October 2004, pp. 2489–2492.

[21] S. J. Young, "The HTK hidden Markov model toolkit: Design and philosophy," Tech. Rep. TR.153, Department of Engineering, Cambridge University, U. K., 1993.

[22] Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 6, pp. 1738–1753, 1990.

[23] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. Int. Conf. Acoust., Speech, and Signal Process.*, 1989, vol. 1, pp. 532–535.