MODEL BASED SPEECH PAUSE DETECTION

Bruce L. McKinley[1] *and Gary H. Whipple*[2]

[1] Signal Processing Consultants
25536 Quits Pond Ct.
South Riding, VA 20152, USA
blmckin@signalprocessing.com

[2] U. S. Department of Defense
9800 Savage Road
Fort George G. Meade, MD 20755, USA

ABSTRACT

This paper presents two new algorithms for robust speech pause detection (SPD) in noise. Our approach was to formulate SPD into a statistical decision theory problem for the optimal detection of noise-only segments, using the framework of model-based speech enhancement (MBSE). The advantages of this approach are that it performs well in high noise conditions, all necessary information is available in MBSE, and no other features are required to be computed. The first algorithm is based on a maximum a posteriori probability (MAP) test and the second is based on a Neyman-Pearson test. These tests are seen to make use of the spectral distance between the input vector and the composite spectral prototypes of the speech and noise models, as well as the probabilistic framework of the hidden Markov model. The algorithms are evaluated and shown to perform well against different types of noise at various SNRs.

1. INTRODUCTION

Accurate detection of noise-only frames in a noisy speech signal, or speech pause detection, is important to many applications such as continuous speech recognition and speech enhancement. Noise can cause severe degradation in recognition tasks, and SPD is required for pre-recognition noise reduction or recognizer model adaptation. SPD is also needed for computing the noise estimate in many speech enhancement systems. Moving-Average Adaptation based on Vector Quantization (MAA-VQ) [2], requires SPD to adapt the noise model in model based speech enhancement [1] for non-stationary noise environments.

Recent algorithms for segmentation of speech do not address noisy conditions and are often iterative schemes not applicable for low-delay communications systems [3]. Previous non-iterative methods, *e.g.* [4]-[5], are significantly degraded by noise or require training under noise conditions identical to those to be encountered by the system. The solution proposed in [6] requires pitch tracking, is limited by the number of class prototypes allowed, and the performance was not validated. A Voiced-Unvoiced classifier for noisy conditions used pitch estimation [7], but pause detection was not included. A variation on that approach used the ratio of unvoiced-to-voiced energies before and after enhancement to reclassify unvoiced segments as pauses [8]. This technique has the disadvantage that the thresholds are empirically derived for a limited number of speech/noise combinations and the thresholds are unpublished. Word boundary detection algorithms developed for isolated word/utterance recognition [9] are not suitable for speech enhancement because the assumption of a single beginning and a single end point does not apply to continuous speech. A technique based on adaptive thresholds attempts to overcome this limitation [10], but it does not perform well for high noise levels unless a two-pass enhancement scheme is used.

The SPD algorithms developed here are optimal detectors based on statistical decision theory, which take advantage of the information computed in model based speech enhancement (MBSE). They are therefore advantageous for use with MBSE, but are also applicable for use in other applications. The detectors are based on maximum *a posteriori* probability (MAP) and Neyman-Pearson (NP) tests which are optimal for the constructed models and received data. They are seen to make use of a spectral distance between the input vector and the composite spectral prototypes of the speech and noise models as well as the probabilistic framework of the hidden Markov Model (HMM), which has proven very useful for many speech processing applications. These SPD algorithms overcome the shortcomings of previous solutions that rapidly degrade in noise, or rely on arbitrary features, heuristic procedures, experimentally derived thresholds, or pitch estimation and tracking. Extensive evaluations are performed, and the detectors are shown to perform well in a variety of noise environments. When used in conjunction with MAA-VQ, they are shown to perform well even in high levels of non-stationary noise.

2. MODEL BASED SPEECH ENHANCEMENT

MBSE estimates clean speech from noisy speech based on parametric mixture HMMs of the clean speech and

noise processes trained via the generalized Lloyd algorithm. The clean speech HMM λ_s is generated off-line (prior to the enhancement process) from extensive training data in order to faithfully represent the set of speech prototypes to be encountered. Additionally, a "silence" state is generated from training data frames with total energy 30 dB below the average of all clean speech frames. The noise HMM λ_v is generated from a period of noise-only data at the beginning of the noisy speech.

In the model based minimum mean-square error approach (MB-MMSE) [1], the clean speech is estimated by filtering the noisy speech with an aggregate filter constructed from the weighted sum of the composite mixture Wiener filters. The composite mixture filters are constructed from the spectral prototypes of the speech and noise in each composite speech/noise model mixture

$$\bar{\alpha}_t = (\bar{\beta}_t, \bar{\gamma}) = \left((\beta_t, \tilde{\beta}_t), (\gamma, \tilde{\gamma})\right), \tag{1}$$

where β_t and $\tilde{\beta}_t$ are the clean speech and noise states, respectively, at time t, and γ and $\tilde{\gamma}$ are the corresponding mixtures for the clean speech and noise states. The spectral prototype of the composite mixture is given by

$$S_{\bar{\alpha}_t}(\theta) = S_{\beta_t,\gamma}(\theta) + S_{\tilde{\beta}_t,\tilde{\gamma}}(\theta) / G_t^2, \tag{2}$$

where $S_{\beta_t,\gamma}$ is the spectral prototype of the clean speech, $S_{\tilde{\beta}_t,\tilde{\gamma}}$ is the spectral prototype of the noise, and G_t^2 is a global gain factor computed to match the input signal power to the clean speech model. The weights for each composite state are simply the conditional probabilities $q(\bar{\alpha}_t \mid \lambda_s, \lambda_v, z_0^t)$ of the composite mixture given the clean speech and noise models and the noisy input data from time 0 to time t. They are computed from the forward-backward formulas [1, eqns 8-10] in conjunction with the probability that the input vector at time t was generated by composite mixture $\bar{\alpha}_t$

$$b(z_t|\bar{\alpha}_t) = \exp\left\{-\frac{K}{2}\ln 2\pi - \frac{1}{2}\sum_{k=0}^{K-1}\ln S_{\bar{\alpha}_t}(\theta) - \frac{1}{2}\sum_{k=0}^{K-1}\frac{|Z_t(\theta)|^2 / G_t^2}{S_{\bar{\alpha}_t}(\theta)}\right\}, \tag{3}$$

where $Z_t(\theta)$ represents the DFT of the noisy speech vector z_t normalized by $K^{1/2}$.

3. OPTIMAL SPEECH PAUSE DETECTORS

The SPD problem can now be formulated into a likelihood ratio test. An (M+1)-hypotheses test is indicated for clean speech silence state β_{sil} corresponding to the null hypothesis H_0 and the other clean speech states 1,...,M corresponding to alternative hypotheses. The SPD-MAP solution results when equal costs are assumed:

$$p(\beta_t = \beta_{sil} \mid z_0^t) \overset{H_0}{>} p(\beta_t = i \mid z_0^t), \quad i = 1, \ldots, M. \tag{4}$$

The posterior probability given the noisy speech data is computed by summing the conditional probability over all composite mixtures containing the clean speech silence state β_{sil}

$$p(\beta_t = \beta_{sil} \mid z_0^t) = \sum_{\gamma=1}^{L_{sil}} \sum_{\tilde{\beta}_t=1}^{\tilde{M}} \sum_{\tilde{\gamma}=1}^{\tilde{L}} q(\beta_t = \beta_{sil}, \gamma, \tilde{\beta}_t, \tilde{\gamma} \mid z_0^t), \tag{5}$$

where L_{sil} is the number of silence mixtures in the clean speech codebook and $\tilde{M}$ and $\tilde{L}$ are the number of states and mixtures in the noise codebook, respectively. Here, for the purpose of clarity, we have not shown the dependence on the clean speech and noise models.

The spectral distance between the gain-normalized input vector and the composite speech/noise spectral prototypes is seen to enter the computation explicitly via the last term in (3). Thus the detector compares the actual second-order statistics of the data versus the models, rather than arbitrary features. It is also superior to the approach outlined in [4], since that technique only had a single template for each class. Additionally, the Markovian nature of the speech prototypes is reflected in the probabilistic framework of the HMM, and enters the computation explicitly through the forward-backward formulas.

The MAP solution may be inappropriate for applications in which there is an unequal cost of making one type of error over another. These include noise model adaptation for MBSE [2], which requires false alarms to be low in order to prevent the noise model from being retrained with speech data (we define false alarms as errors in which the SPD classifies a given frame as a pause when it in fact contains speech). In this case, it is desirable to trade missed detections for decreased false alarm rate. A Neyman-Pearson test (SPD-NP) can be formulated by considering a single alternative hypothesis. Since

$$p(\beta_t \neq \beta_{sil} \mid z_0^t) = 1 - p(\beta_t = \beta_{sil} \mid z_0^t), \tag{6}$$

the likelihood ratio test reduces to comparison of the probability of clean speech silence to a threshold

$$p(\beta_t = \beta_{sil} \mid z_0^t) \overset{H_0}{>} \eta, \tag{7}$$

where the threshold η is chosen to yield a given maximum false alarm probability.

4. EVALUATIONS

The SPD-MAP and SPD-NP algorithms were evaluated for additive white Gaussian (AWGN) and pink noise, and a slightly non-stationary Lynx helicopter noise environment (from the NOISEX-92 set, RSG.10 NOISEROM-0), at SNRs from -5 to 30 dB. The MB-MMSE

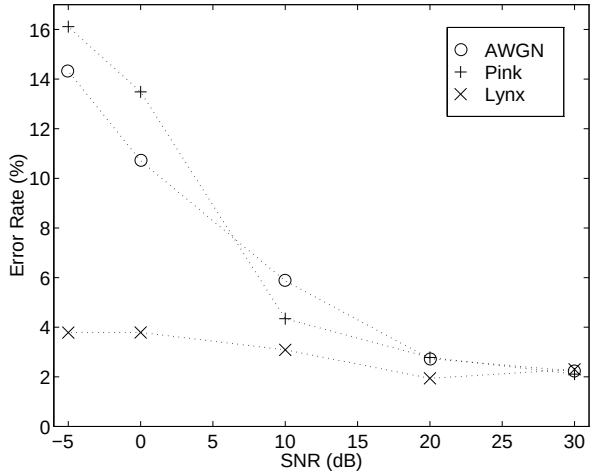

Figure 1. Performance of SPD-MAP.

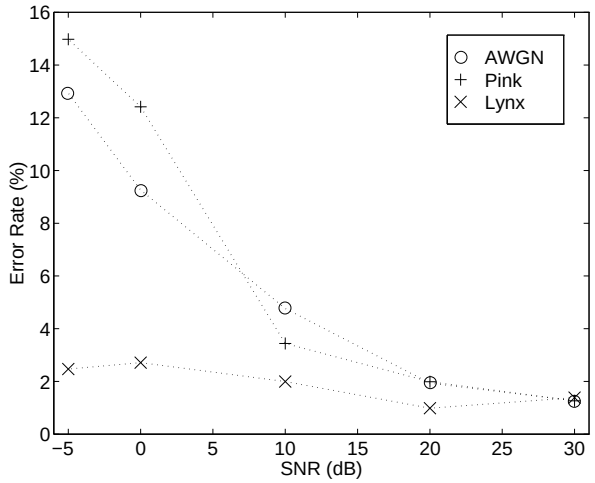

Figure 2. Performance of SPD-MAP, excluding transition frames.

system used an 8 state × 32 mixture (plus a silence state with 32 mixtures) clean speech codebook of order 10 trained on 300 sentences from the TIMIT database, and an 8×1 noise codebook of order 50. Eighteen sentences from six different speakers in the TIMIT database were used in testing. The test set consisted of a total of 2403 frames of data, of which 651 were manually identified as pauses. The test speakers and sentences were different from those used in training. The resulting SPD is smoothed using a median filter of length 5.

The SPD-MAP algorithm was seen to be effective over all noise types and SNRs examined, as shown in Figure 1. An error rate less than 2.3% is achieved for input SNRs of 30 dB. This surpasses the results presented in [3] (2.4% error rate under no noise, rapidly degraded in noise) and [4] (2.5% error rate under quiet telephone line conditions). As noted in [4], many of the errors are made in transition frames, which can contain speech as well as a partial pause. If desired, these false alarm errors can be

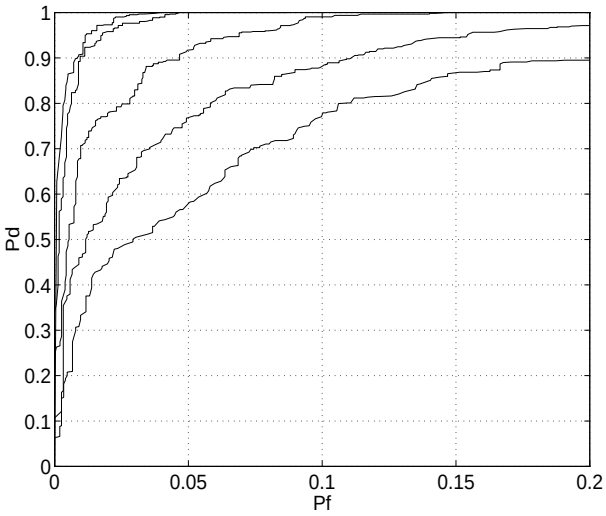

Figure 3. Receiver Operating Curves of SPD-NP in AWGN at input SNRs of 30, 20, 10, 0 and -5 dB.

TABLE I
PERFORMANCE OF SPD-NP IN AWGN, THRESHOLD $\eta = 0.4$.

Input SNR (dB)	Detection Errors: Silence (651)	Speech (1752)	Total (2403)	Probability of Error
30	43	14	57	0.024
20	51	17	68	0.028
10	74	73	147	0.061
0	71	183	254	0.106
-5	60	321	381	0.159

prevented by discarding the first and last frames of each identified pause to achieve the results presented in Figure 2, where error rates less than 1.4% are achieved, with graceful degradation at lower SNRs.

The Neyman-Pearson solution is an alternative method of reducing false alarms. The receiver operating curves (ROCs) for SPD-NP are shown in Figure 3. There it is seen that good detection results can be achieved while maintaining low probability of false alarm, P_F. Table I gives the results for SPD-NP in AWGN using a threshold η=0.4. The detection errors are tabulated for frames identified as "Silence" or "Speech." Errors in the Silence column indicate false alarms, whereas error in the Speech column indicate missed detections. The results show good detection performance while maintaining P_F at or below 1.1% for all SNRs tested.

A final evaluation was performed using F16 jet engine noise which exhibits significant non-stationarity in power and frequency content, making SPD difficult. A 3-dimensional spectrogram of this noise environment is shown in Figure 4. This type of environment exemplifies the variations in noise characteristics found in common applications, for which typical SPD schemes will fail.

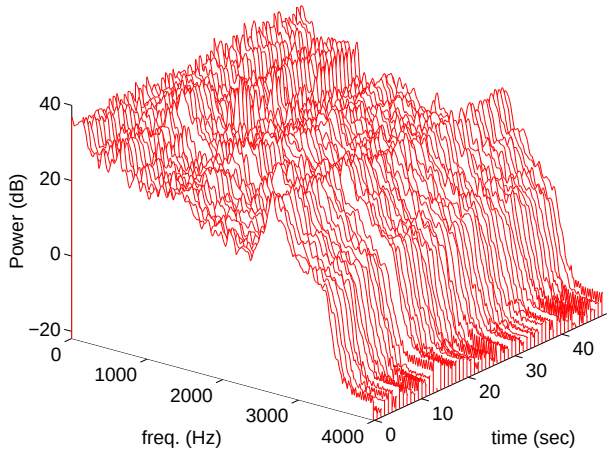

Figure 4. Spectrogram of F16 noise from NOISEX-92.

However, when used in conjunction with noise model adaptation (MAA-VQ), the SPD-NP detector with threshold η=0.4 continues to reliably detect pauses even at the lowest SNRs. These results are shown in Figure 5.

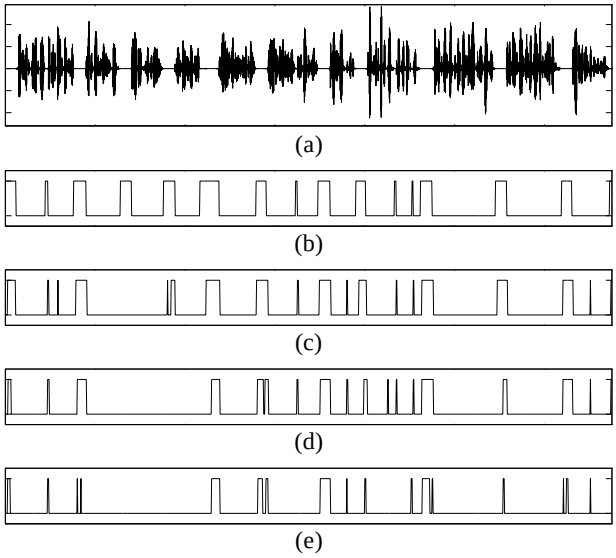

Figure 5. Operation of SPD-NP (η=0.4) in non-stationary F16 noise. (a) 11 sentences of speech with no noise; (b) manually identified pauses; (c)-(e) SPD-NP at input SNRs of 10, 0 and -5 dB, respectively.

5. CONCLUSIONS

The speech pause detection problem was formulated into a decision theory framework based on models of the speech and noise processes, and optimal MAP and NP detectors were developed. The detectors were shown to perform well even at low SNRs. The method, when used in conjunction with MAA-VQ, was also shown to be effective for non-stationary environments. The same detection framework might be used for a more generic segmentation of speech into other categories such as voiced/unvoiced/pause by using additional information extracted from the clean speech model training procedure.

Additionally, a clean speech HMM containing a silence state trained with data specifically identified as non-speech can yield more accurate results than the one trained using the energy threshold described in [1] and used in this work.

REFERENCES

[1] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models." IEEE Trans. Signal Processing, vol. 40, no. 4, pp. 725-735, April 1992.

[2] B. L. McKinley and G. H. Whipple, "Noise model adaptation in model-based speech enhancement," Proc. 1996 IEEE ICASSP (Atlanta), May 1996, pp. 633-636.

[3] E. Vidal and A. Marzal, "A review and new approaches for automatic segmentation of speech signals," Signal Processing V, L. Torres, E. Masgrau and M. A. Lagunas (eds). Elsevier Science Publishers B V. 1990.

[4] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 3, pp. 201-212, June 1976.

[5] L. R. Rabiner and M. R. Sambur, "Application of an LPC distance measure to the Voiced-Unvoiced-Silence detection problem," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, no. 4, pp. 338-343, Aug. 1977.

[6] R. J. McAuley, "Optimum classification of voiced speech, unvoiced speech and silence in the presence of noise and interference," MIT Lincoln Lab., Lexington, MA, Tech. Note 1976-7, June 1976.

[7] D. A. Krubsack and R. J. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, pp. 319-328, Feb. 1991.

[8] H. Sheikhzadeh, R. L. Brennan, and H. Sameti, "Real-time implementation of HMM-based MMSE algorithm for speech enhancement in hearing aid applications," *Proc. 1995 IEEE ICASSP* (Detroit), May 1995, pp. 808-811.

[9] J.-C. Junqua, B. Mak and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 3, pp. 406-412, July 1994.

[10] J. H. L. Hansen, "A new speech enhancement algorithm employing acoustic endpoint detection and morphological based spectral constraints," *Proc. 1991 IEEE ICASSP* (Toronto), May 1991, pp. 901-904.