DySANA: Dynamic Speech and Noise Adaptation for Voice Activity Detection

Ron J. Weiss[1], *Trausti Kristjansson*[2]

[1]Columbia University, New York, NY, USA
[2]Google Inc., New York, NY, USA
ronw@ee.columbia.edu, trausti@google.com

Abstract

We describe a method of simultaneusly tracking noise and speech levels for signal-to-noise ratio adaptive speech endpoint detection. The method is based on the Kalman filter framework with switching observations and uses a dynamic distribution that 1) limits the rate of change of these levels 2) enforces a range on the values for the two levels and 3) enforces a ratio between the noise and the signal levels. We call this a *Lombard dynamic distribution* since it encodes the expectation that a speaker will increase his or her vocal intensity in noise. The method also employs a state transition matrix which encodes a prior on the states and provides a continuity constraint. The new method provides 46.1% relative improvement in WER over a baseline GMM based endpointer at 20 dB SNR.

Index Terms: voice activity detection, endpointing, Kalman filter, Lombard effect

1. Introduction

We consider the problem of noise robust speech detection in the context of a real time ASR dialog system. A separate module, called an *endpointer* is commonly used for this purpose. A dialog system requires the determination of both the start and end times of speech with low latency and preferably with low computational requirements. By accurately identifying the speech endpoints, the recognition accuracy is increased and computational requirments are reduced. In addition, an endpointer identifies if a user has spoken during prompt so that the system can stop the prompt. The endpointer also signals the end of speech, allowing the recognizer to determine if a valid recognition has been found.

A simple method for speech endpoint detection is to compute the energy of a signal and assume that speech is present if the energy exceeds a threshold. Since the noise and speech levels vary by environment, strategies for adapting the threshold are employed [1]. Another common endpointing strategy is to use a speech/noise classifier, such as a Gaussian Mixture Model (GMM) based classifier. The approach described in this paper can be seen as a combination of these strategies.

Rennie et al. [2] and Fujimoto et al. [3] both tackle the problem of tracking the noise level for denoising and voice activity detection, respectively. They describe related paradigms that employ parallel model combination in the log spectrum or log-Mel spectrum domains. They both employ a first order continuous dynamic process for tracking a noise parameter. Fujimoto et al. [3] treat the noise as a variable and track the noise variable directly whereas Rennie et al. treat the noise as a random variable and track the noise *level*. Defining the dynamics on the noise level has benefits related to seperately controlling the rate at which the parameter is allowed to change while allowing the observation to infuence the rate of change.

Unlike the above methods, we work in the cepstral domain and employ a model with switching observation distributions. In other words, we make the approximation that the observation is explained either by the speech model or the noise model. In the current work, we decompose the observed features into a set that is invariant to the signal level and a set that depend on it and needs to be adapted.

Since the endpointer needs to make decisions very quickly after the speech event arrives, and after having observed a very limited amount of data, this strategy allows the endpointer to rely on the invariant features until good estimates have been found for the remaining model components. We assume that only the model components that relate to the gain of the signal and the gain of the noise are dependent on the environment and channel. In the cepstral domain, only a single component, i.e. the C_0 component, is dependent on the gain.

2. Noisy speech signal model

We model the observed signal at frame t, $\mathbf{y}^t$, using standard Mel frequency cepstral coefficient (MFCC) speech features. The key property of the MFCC representation is that the majority of the feature dimensions are independent of the signal energy. This implicit factoring of the representation allows us to efficiently track the instantaneous signal and noise levels while still relying on the spectral shape information in making the endpointing decision.

The 0th MFCC is a function of the signal energy level while the remaining dimensions only capture information about the signal's spectral shape. Therefore, we define $\boldsymbol{\ell} = [1 \quad \mathbf{0}]^T$ and partition $\mathbf{y}^t$ into the level-dependent term $\boldsymbol{\ell}^T\mathbf{y}^t = y_0^t$ and the spectral shape term $\mathbf{y}_{1:D}^t$.

The DySANA signal model operates on these observations using a Kalman filter with switching observations [4]. At each frame, the observation is explained by either a speech Gaussian mixture model (GMM), or a silence/noise model. We define the random variable s^t to denote this decision for frame t. $s^t = 1$ if speech is present in frame t, and $s^t = 0$ otherwise. This switching behavior between different observation distributions is what makes the DySANA model a switching Kalman filter. Finally, as in [3], we smooth s^t using a simple hidden Markov model (HMM) to prevent the endpointer from switching too rapidly between the speech and noise states.

At each frame we track the speech and noise levels relative to their respective models. Each model is parametrized by a set of Gaussian means and diagonal covariance matrices: $\{\boldsymbol{\mu}_{x,c_x}, \Sigma_{x,c_x}\}_{c_x \in 1..N_x}$ for speech and

$\{\boldsymbol{\mu}_{n,c_n}, \Sigma_{n,c_n}\}_{c_n \in 1..N_n}$ for noise. We define the speech "gain" g_x^t as the difference between the observed signal level y_0^t and the mean level of the MAP component for frame t in the speech model, $\mu_{x,\hat{c}_x,0}^t$. The noise gain, g_n^t, is defined analogously. The combined speech and noise gains $\mathbf{g} = [g_x^t \quad g_n^t]^T$ comprise the Kalman filter state space.

The joint likelihood of the parameters for frame t given all previous observations can be factored as follows:

$$P(\mathbf{y}^t, c^t, s^t, \mathbf{g}^t | Y^{t-1}) = P(\mathbf{y}^t | c^t, \mathbf{g}^t, s^t) P(c^t | s^t) P(s^t | Y^{t-1}) P(\mathbf{g}^t | Y^{t-1}) \quad (1)$$

where Y^t denotes the set of observations up to and including time t. The first term is the likelihood of the current frame under the given GMM component c^t and the current estimates of the dynamic parameters s^t and $\mathbf{g}^t$. $P(c^t|s^t)$ is simply the prior over the mixture components of the speech and noise models. $P(s^t|Y^{t-1})$ models the dynamics of the speech decision variable and $P(g^t|Y^{t-1})$ models the Kalman filter dynamics of the gain parameters. We model this using a Gaussian distribution:

$$P(\mathbf{g}^t | Y^{t-1}) = \mathcal{N}(\mathbf{g}^t; \boldsymbol{\mu}_{\mathbf{g}^t}, \Sigma_{\mathbf{g}^t}) \quad (2)$$

$$= \mathcal{N}\left(\begin{bmatrix} g_x^t \\ g_n^t \end{bmatrix}; \begin{bmatrix} \mu_{g_x^t} \\ \mu_{g_n^t} \end{bmatrix}, \begin{bmatrix} \sigma_{g_x^t} & \sigma_{g_{xn}^t} \\ \sigma_{g_{xn}^t} & \sigma_{g_n^t} \end{bmatrix} \right) \quad (3)$$

The endpointing decision is based on the conditional posterior of s^t:

$$P(s^t{=}1 | Y^t) \propto P(s^t{=}1 | Y^{t-1}) P(\mathbf{y}^t | s^t{=}1, Y^{t-1}) \quad (4)$$

$$= P(s^t{=}1 | Y^{t-1}) \sum_{c_x^t} P(c_x^t | s^t{=}1) z_x(\mathbf{y}^t) \quad (5)$$

where

$$z_x(\mathbf{y}^t) = \mathcal{N}(\mathbf{y}^t; \boldsymbol{\mu}_{x,c_x^t} + \boldsymbol{\ell}\mu_{g_x^t}, \Sigma_{x,c_x^t} + \boldsymbol{\ell}\boldsymbol{\ell}^T \sigma_{g_x^t}) \quad (6)$$

3. Model dynamics

In order to compute the speech posterior for successive frames, it is necessary to propagate the distribution of the dynamic parameters s and $\mathbf{g}$. As shown above, the posterior distribution over s^{t+1} depends on the statistics of the dynamic parameters from the previous frame, s^t and $\mathbf{g}^t$, given the observed signal up to frame t.

The dynamics of s^t are identical to those of the HMM forward algorithm. Therefore, the conditional prior of s^{t+1} can be found simply by multiplying the posterior of s^t by the speech/nonspeech transition matrix.

$$P(s^{t+1} | Y^t) = \sum_{s^t} P(s^t | Y^t) P(s^{t+1} | s^t) \quad (7)$$

The conditional gain prior for frame $t+1$ has the analogous form for a continuous random variable.

$$P(\mathbf{g}^{t+1} | Y^t) = \int_{\mathbf{g}^t} P(\mathbf{g}^t | Y^t) P(\mathbf{g}^{t+1} | \mathbf{g}^t) \quad (8)$$

Proper selection of the gain transition distribution $P(\mathbf{g}^{t+1}|\mathbf{g}^t)$ is key to achieving the desired performance. A good first order approximation for the dynamic distribution is a random walk model which allows the state variable for time $t+1$ to move away from the estimate for time t in any direction:

$$P(\mathbf{g}^{t+1} | \mathbf{g}^t) = \mathcal{N}(\mathbf{g}^{t+1}; \mathbf{g}^t, \Sigma_{RW}) \quad (9)$$

However, such a dynamic distribution is problematic for use with Kalman filters with switched observations. Recall that only the speech or noise gain are observed in any given frame. Therefore if such dynamics are used, the variance for the unobserved variable can grow without bounds during long periods of silence or speech. Furthermore, there are no constraints on the limits of $\mathbf{g}^t$. So, for example, it is possible for the gain estimates to predict a very high noise gain which may result in speech frames being misclassified as nonspeech.

To control these problems, we introduce constraints on the dynamic distribution in the form of a prior-like factor over $\mathbf{g}^{t+1}$. This leads to the *Lombard dynamic distribution*

$$P(\mathbf{g}^{t+1} | \mathbf{g}^t) \propto \mathcal{N}(\mathbf{g}^{t+1}; \mathbf{g}^t, \Sigma_{RW}) \mathcal{N}(\mathbf{g}^{t+1}; \boldsymbol{\mu}_{SNR}, \Sigma_{SNR}). \quad (10)$$

This distribution has the effect of introducing an SNR coupling between the two signals. This effect is intuitively appealing as it allows the model to assume that the speech level will be higher if a high noise level has been observed. This enables it to capture the Lombard effect[1]. Notice in figure 1 that during the initial noise from 0.0 sec to about 1.5 sec, the speech gain follows the noise gain, even though only noise is observed. Additionally, this constraint prevents $\mathbf{g}^t$ from straying too far from the prior level, and prevents it's variance from growing too large. An interesting aspect of the dynamic distribution is that it allows one to tune the performance over a range of ROC curves e.g. between the DySANA-p and the DySANA curves in figure 2.

As shown in section 2, the posterior of $\mathbf{g}^t$ given Y^t in equation 8 is a mixture of Gaussians. This implies that the full distribution over $\mathbf{g}^{t+1}$ has a distribution with $N_x + N_n$ modes. Propagating such a complex distribution can be expensive. Instead, as in [2], we approximate it with a single Gaussian at the most probable mode of the full distribution. This occurs at the maximum a posteriori settings of c^t and s^t, $\hat{c}^t$ and $\hat{s}^t$. For example, if the MAP setting has $\hat{s}^t = 1$,

$$P(\mathbf{g}^{t+1} | Y^t) \approx \mathcal{N}(\mathbf{g}^{t+1}; \boldsymbol{\mu}_{\mathbf{g}^{t+1}}, \Sigma_{\mathbf{g}^{t+1}}) \quad (11)$$

where

$$\boldsymbol{\mu}_{\mathbf{g}^{t+1}} = W\boldsymbol{\mu}_{\ell p} + (I - W)\boldsymbol{\mu}_{SNR} \quad (12)$$

$$\Sigma_{\mathbf{g}^{t+1}} = W(\Sigma_{RW} + \Sigma_{\ell p}) \quad (13)$$

$$\boldsymbol{\mu}_{\ell p} = \Sigma_{\ell p} \left(\Sigma_{\mathbf{g}^t}^{-1} \boldsymbol{\mu}_{\mathbf{g}^t} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{y_0^t - \mu_{x,\hat{c}_x^t,0}}{\sigma_{x,\hat{c}_x^t,0}} \right) \quad (14)$$

$$\Sigma_{\ell p} = \left(\Sigma_{\mathbf{g}^t}^{-1} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \frac{1}{\sigma_{x,\hat{c}_x^t,0}} \right)^{-1} \quad (15)$$

$$W = \Sigma_{SNR}(\Sigma_{SNR} + \Sigma_{RW} + \Sigma_{\ell p})^{-1} \quad (16)$$

The propagated mean $\boldsymbol{\mu}_{\mathbf{g}^{t+1}}$ is a weighted combination of the conditional prior gain from the previous observation, $\boldsymbol{\mu}_{\mathbf{g}^t}$, the SNR gain prior, $\boldsymbol{\mu}_{SNR}$, and the gain estimate based on the observation. Since observing speech gives no new information about the instantaneous noise gain, $\boldsymbol{\mu}_{\ell p}$ and $\Sigma_{\ell p}$ reduce to the previous values for the noise gain. This forces the noise gain to drift toward the prior $\boldsymbol{\mu}_{SNR}$ during a long sequence of speech observations. Since Σ_{SNR} is a full matrix, W is also full. This allows the observation of speech to influence the update for the

[1] The Lombard effect is the tendency to increase one's vocal intensity in noise.

noise model. As described above, the variance of gain of the unobserved model increases at each time step, but its growth is bounded by the dynamic distribution.

The derivation for the case where $\hat{s}^t = 0$ is similar, except $\boldsymbol{\mu}_{\ell p}$ and $\Sigma_{\ell p}$ are as follows.

$$\boldsymbol{\mu}_{\ell p} = \Sigma_{\ell p}\left(\Sigma_{\mathbf{g}^t}^{-1}\boldsymbol{\mu}_{\mathbf{g}^t} + \begin{bmatrix}0\\1\end{bmatrix}\frac{y_0^t - \mu_{n,\hat{c}_n^t,0}}{\sigma_{n,\hat{c}_n^t,0}}\right) \quad (17)$$

$$\Sigma_{\ell p} = \left(\Sigma_{\mathbf{g}^t}^{-1} + \begin{bmatrix}0 & 0\\0 & 1\end{bmatrix}\frac{1}{\sigma_{n,\hat{c}_n^t,0}}\right)^{-1} \quad (18)$$

An example of DySANA speech and noise gain tracking can be seen in figure 1. The GMM endpointer makes many false accept errors during the noise burst at about 3.0 seconds while DySANA is able to adapt to the increased noise level and correctly classify the noisy frames. As shown in the bottom panel, the gain of the observed model tracks the observation, but is falls back to the prior level (0) when unobserved. Because we use a full covariance matrix for Σ_{SNR}, the estimate for the unobserved model sometimes changes, tracking that of the observed model. Finally, we note that the variance of unobserved model increases with the number of frames since previous observation as on the transition from noise to speech at 1.5 seconds. Again, this implies decreased reliance on the level-dependent features of the unobserved model for the endpointing decision, instead backing off to spectral shape features.

4. Experiments

To evaluate the performance of the proposed endpointing algorithm, we assembled a dataset of noisy speech signals based on the DNA database of car noise recordings [2] and the AURORA2 framework for noisy speech recognition [5]. Clean speech signals from the AURORA2 test set were mixed with car noise from the DNA database at signal-to-noise ratios varying between 0 and 20 dB in increments of 5 dB. The noisy utterances were then passed through the AMR speech coder/decoder chain [1] to simulate the processing applied to cell phone signals. Finally, the resulting noisy signals were broken up into utterances designed to mimic interactions with a dialog system. 8% of the resulting utterances were composed of 3 seconds segments containing no speech at all. The remaining utterances were composed of 3 seconds of noise followed by the noisy speech utterance, followed by an additional 2 seconds of noise. The final dataset consisted of a total of about 100 minutes of data, split 66%, 33% development and testing respectively. The results reported in section 5 are over the test set only.

We compare the proposed endpointer to a simple endpointer based on an adaptive energy threshold (Energy), the ETSI AFE endpointer described in [6] (ETSI AFE), a baseline statistical model endpointer based on an unadapted GMM classifier (GMM), and the switching Kalman filter endpointer described in [3] (SKF) that tracks the noise process underlying the speech signal. Finally, we evaluate a few variants of the proposed algorithm: the full DySANA endpointer described in the previous section, the DySANA endpointer without HMM smoothing (DySANA-h), the DySANA endpointer without prior contraints (i.e. only using random walk dynamics) (DySANA-p), and DySANA-p-h using neither. All statistical model based systems used the same 32 component speech and nonspeech models trained on data collected from the Goog411 dialog system.

The final VAD decision for the statistical model-based endpointers is made by first thresholding the posterior probability that a given frame is dominated by speech (i.e. $P(s^t)$) and then feeding the resulting binary decision to a simple finite state machine similar to that used in [6], designed to smooth the output.

System	0	5	10	15	20	Clean
No VAD	106.5	97.8	81.7	70.1	63.5	4.8
ETSI AFE	93.7	87.5	78.7	59.6	57.5	7.9
Energy	106.5	96.5	76.7	56.4	30.0	3.8
GMM	79.7	63.4	35.2	22.2	11.5	3.8
SKF-h	75.5	48.6	27.4	17.2	9.0	3.8
SKF	78.8	51.6	27.8	17.2	8.7	3.9
DySANA-p-h	66.7	48.1	26.3	14.5	6.5	3.3
DySANA-p	**64.6**	47.3	27.7	14.6	7.8	3.4
DySANA-h	67.5	**44.0**	24.7	15.9	7.5	3.6
DySANA	74.2	46.8	**23.9**	**13.5**	**6.2**	**3.3**

Table 1: Word error rate of different endpointers as a function of SNR. Note that data was distorted by AMR encoding and decoding to match cell phone data.

The testing subset of the data was used to determine the best parameter settings for the different algorithms. We performed a grid search over the adaptation parameters and decision threshold for each system and chose parameters that maximized the average word error rate across all SNRs. For the SKF endpointer, the observation variance was set to 1.0. For DySANA, $\boldsymbol{\mu}_{SNR} = \mathbf{0}$, $\Sigma_{SNR} = \begin{bmatrix}100 & 10\\10 & 40\end{bmatrix}$, and $\Sigma_{RW} = \begin{bmatrix}10 & 0\\0 & 2.5\end{bmatrix}$ were found to work best. Finally, the speech/nonspeech transition matrix for all endpointers was chosen such that the stationary distribution had a speech prior of 0.23, which matched that of the test set.

5. Results

Table 1 shows the recognition performance using the different algorithms described above. The speech recognzier used was the multicondition Aurora2 HTK recognizer trained over AMR coded speech. The statistical model based endpointers significantly outperform the ETSI-AFE and energy-based systems in all noise conditions. The GMM endpointer performs very poorly under the noisiest conditions where the model used was an extremely poor fit for the data. The adaptive algorithms all improve on this baseline, with the full DySANA system performing best under all but the noisiest conditions where errors result from the SNR prior distribution discouraging the system from tracking extremely high noise levels. The systems that utilize HMM smoothing tend to perform better than the same system without smoothing under less noisy conditions, however at low SNRs the smoothing sometimes reinforces erroneous classifications, resulting in reduced performance. In 0 dB conditions the DySANA variant that only uses random-walk dynamics performs best, but as the SNR increases it does not perform as well.

The ROC performance of statistical model based algorithms are shown in figure 2. Again, the DySANA endpointer performs best in general. However, when removing the gain prior distribution from the Kalman filter dynamics DySANA becomes more skewed towards false reject errors. This is a result of the behavior described in section 3 where the noise gain tracks too high, resulting in many misclassification errors of speech as nonspeech. While this problem does not appear very severe at the frame level, it is clearly a significant issue at the whole utterance level, explaining the decreased recognition performance of this

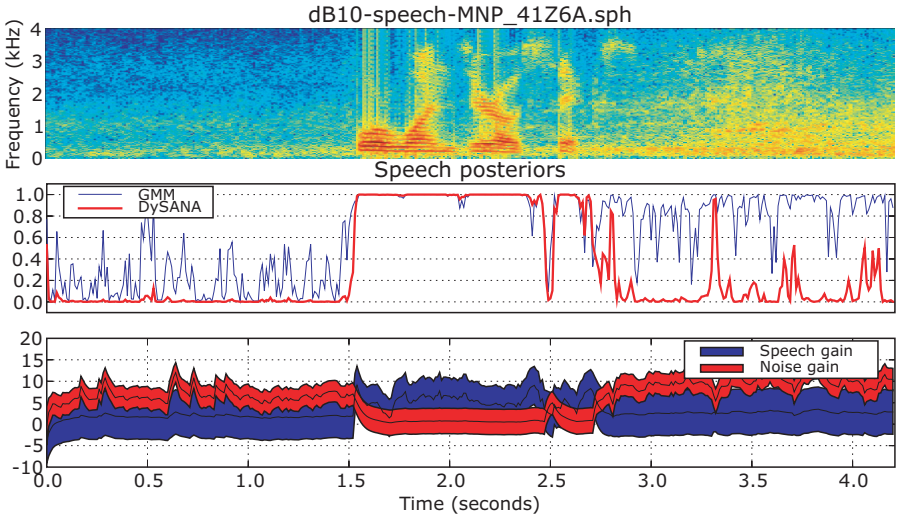
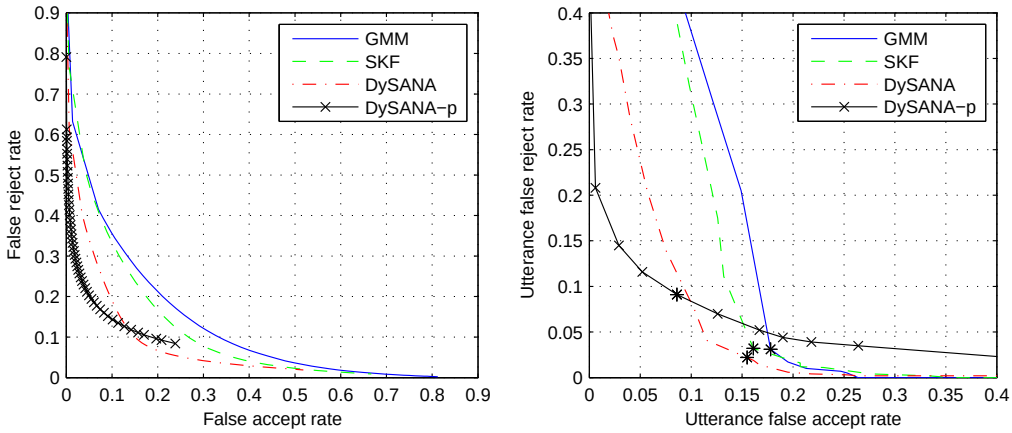


Figure 1: Example of DySANA speech and noise level adaptation. The middle panel shows the speech posterior $P(s^t|Y^t)$ of the signal displayed in the top panel using a baseline GMM classifier (GMM) and using the DySANA endpointer. The bottom panel shows the DySANA VAD's estimate of the speech and noise gain. The width of each gain track denotes the associated variance.



(a) Frame-level ROC

(b) Utterance-level ROC

Figure 2: ROC curves averaged across all noise conditions. (a) shows the relationship between false accept and false rejects at the frame level and (b) shows the utterance level. The stars in (b) denote the operating points of the endpointers corresponding to the results in table 1.

system in table 1.

6. Conclusion

We have presented a method for signal-to-noise ratio adaptive speech endpoint detection based on a switching Kalman filter framework for tracking the instantaneous speech and noise levels. A key to this method's success is a dynamic distribution that limits the range of values that the noise and speech models can take and introduces a coupling between the levels. When applied to speech corrupted by car noise, the proposed method shows significant improvement over an unadapted GMM classifier based endpointer.

7. References

[1] "ANSI-C code for the adaptive multi rate (AMR) speech codec," June 2007, 3GPP standard document. 3GPP TS 26.073 V7.0.0.

[2] S. Rennie, T. Kristjansson, P. Olsen, and R. Gopinath, "Dynamic noise adaptation," in *Proceedings of ICASSP*, 2006.

[3] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching Kalman filter," in *Proceedings of Interspeech*, 2007, pp. 2933– 2936.

[4] K. Murphy, "Switching Kalman filters," U. C. Berkeley, Tech. Rep., 1998.

[5] H.-G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR-2000*, 2000, pp. 181–188.

[6] "Speech processing, transmission and quality aspects (STQ), distributed speech recognition, advanced front-end feature extraction algorithm, compression algorithms," 2007, eTSI standard document. ETSI ES 202 050 V1.1.5.