

Speech Activity Detection Fusing Acoustic Phonetic and Energy Features

Etienne Marcheret, Karthik Visweswariah, Gerasimos Potamianos

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

{etiennem, kv1, gpotam}@us.ibm.com

Abstract

With the wider deployment of automatic speech recognition (ASR) systems, the importance of robust speech activity detection has been elevated both as a means of reducing bandwidth in client/server ASR and for overall system stability from barge-in through the recognition process. In this paper we investigate a novel technique for speech activity detection, that we have found to be effective in handling non-stationary noise events without negatively impacting the recognition process. This technique is based on combining acoustic phonetic likelihood based features with energy features extracted from the signal waveform. Reported results on two speech activity detection tasks demonstrate that the proposed method outperforms techniques which rely solely on acoustic or energy features.

1. Introduction

Speech activity detection has long been an important issue as a front end step to the ASR process. Its significance ranges (although not limited to) from bandwidth usage in the client/server ASR paradigm, to stable prompt control during barge-in operation, with positive ASR impact in terms of both CPU and accuracy. The impact on CPU is straightforward since the decoder is not required to operate on non-speech segments. The impact on accuracy is somewhat more complicated, since, on the one hand, elimination of silence segments decreases the decoder insertion error rate simply by restricting the amount of audio reaching it, but, on the other hand, silence segments can generate observations useful to on-line adaptation.

Not surprisingly, speech activity detection has attracted significant interest in the ASR literature. Most techniques are based on features extracted from the acoustic signal, ranging from plain energy [1] to frequency-based representations of speech [2]–[4]. The selected features are subsequently used in speech/silence classification, ranging from adaptive thresholding to linear discriminants, regression trees, distance measures, and Gaussian mixture model based classifiers. In general, energy-based speech detection is computationally efficient and simple to implement, but it lacks robustness to noise. Performance can be improved by using adaptive thresholds or appropriate filtering of the energy estimates [1, 5], however addressing non-stationary noise effectively remains difficult. Most often, frequency-based speech features, such as mel-frequency cepstral coefficients (MFCCs), are required to achieve improved robustness to noise. In this paper, we propose to employ such features indirectly, through the acoustic model that is assumed to generate them. The resulting acoustic phonetic features are extracted based on the phonetic class conditional MFCC observation vector likelihoods by the acoustic model, and are used to augment baseline energy based features. The two types of features are fused and subsequently considered for speech/silence detection using a Gaussian mixture classifier. The proposed algorithm is tested in two domains, a narrowband telephony task, and on

wideband, far-field acoustic data, collected as part of the CHIL project [6].

The rest of the paper is divided into two main parts: Section 2 that discusses the proposed algorithm, and Section 3, where experimental results are presented. Finally, Section 4 provides a short summary.

2. Fused Acoustic Phonetic and Energy Feature Speech Activity Detection

The speech activity detection system operates on two types of features: Energy based ones, generated directly from the waveform, and acoustic phonetic features, defined from observations generated by the ASR acoustic model. The two feature sets are combined, and are subsequently fed to a Gaussian mixture model (GMM) classifier. These steps are presented next.

2.1. Energy Based Features

The energy based feature space is defined by a five-dimensional vector, the components of which are based on the bandpass filtered acoustic waveform. In particular, for the telephony task the passband is from 200 to 3800 Hz, whereas the wideband task uses the entire [0, 11] kHz passband. Letting $y[i]$ denote the bandpass-filtered waveform at sample time i , the estimated short time energy $e(t)$ for a window of length N is given by

$$e(t) = 10 \log \left(\frac{1}{N} \sum_{i=1}^N y[i]^2 \right), \quad (1)$$

measured in dB. In (1), index t is discrete and determined by the observation frame rate, set in this work to be 10 msec. Given $e(t)$, we generate filtered observations of it, based on

$$rms(t) = 10^{scale \times e(t)}. \quad (2)$$

In (2), $rms(t)$ is defined as a linear energy scaled for the expected number of bits of resolution. The scaling constant is given by $scale = contrast/scaleMax$, where the $contrast$ provides a level of sensitivity, generally set within [3.5, 4.5], and $scaleMax$ is the maximum possible value of $e(t)$, for example 90.3 dB for a 16-bit signed linear PCM signal.

Based on the instantaneous $rms(t)$ value, we can obtain “low”, “mid”, and “high” energy tracks, defined as

$$lt(t) = (1 - \alpha_{l,t}) \times lt(t-1) + \alpha_{l,t} \times rms(t) \quad (3a)$$

$$mt(t) = (1 - \alpha_m) \times mt(t-1) + \alpha_m \times rms(t) \quad (3b)$$

$$ht(t) = (1 - \alpha_{h,t}) \times ht(t-1) + \alpha_{h,t} \times rms(t) \quad (3c)$$

respectively. For the mid-track $mt(t)$, time constant α_m is fixed, set in this work to 0.1. Therefore $mt(t)$ is the lowpass filtered $rms(t)$. The remaining low and high track time constants $\alpha_{l,t}$ and $\alpha_{h,t}$ are functions of the instantaneous $rms(t)$,

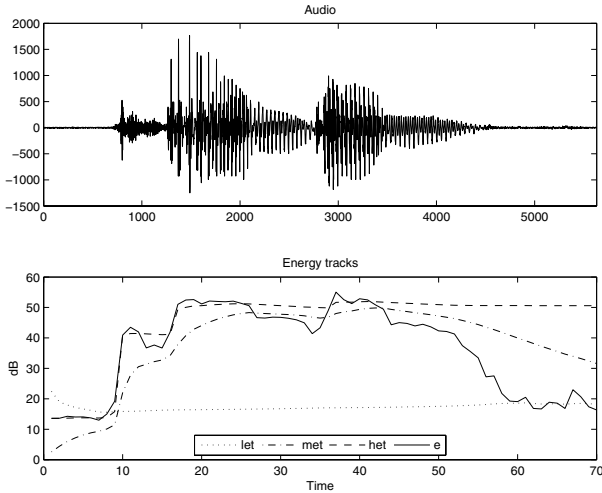


Figure 1: Audio waveform and corresponding energy tracks

designed so that rapid changes in energy will cause abrupt tracking by $lt(t)$ and $ht(t)$, respectively. They are given by

$$\alpha_{l,t} = \left(\frac{lt(t-1)}{rms(t)} \right)^2 \quad \text{and} \quad \alpha_{h,t} = \left(\frac{rms(t)}{ht(t-1)} \right)^2,$$

thus resulting in increasing $\alpha_{l,t}$ for decreasing $rms(t)$, and increasing $\alpha_{h,t}$ for increasing $rms(t)$.

Next, from (3), we form three equivalent low, mid, and high energy representations, given by

$$\begin{aligned} let(t) &= \frac{\log(lt(t))}{scale}, & met(t) &= \frac{\log(mt(t))}{scale}, \\ \text{and} & & het(t) &= \frac{\log(ht(t))}{scale}. \end{aligned} \quad (4)$$

From (4), we can also obtain the mid energy to low energy track relationship as

$$m2l(t) = met(t) - let(t). \quad (5)$$

By combining (1), (4), and (5) we obtain a five-dimensional energy feature vector at frame t , as

$$v_e(t) = [e(t) \ let(t) \ met(t) \ het(t) \ m2l(t)]. \quad (6)$$

A purely energy based speech activity detector based on these observations where dynamic speech/silence thresholds are employed can be found in [5].

Fig. 1 shows the bandpassed filtered energy (1) and the corresponding energy tracks let , met , and het defined in (4). We observe that the low and high energy tracks are intended to lock onto the floor and speech signal levels respectively, while the mid track is a lowpass filtered energy track.

2.2. Acoustic Phonetic Features

The acoustic phonetic feature space employed for speech activity detection is derived from the acoustic model used for ASR. The acoustic model is generated from partitioning the acoustic space by context-dependent phonemes with the context defined in this work as plus and minus five phonemes, cross-word to the left only. The context-dependent phoneme observation generation process is modeled as a GMM within the hidden Markov model (HMM) framework, and in typical large-vocabulary ASR

systems, this leads to more than 40k Gaussian mixture components. Calculating all HMM state likelihoods from all Gaussians at each frame would preclude real-time operation. Therefore, we define a hierarchical structure for the Gaussians, where it is assumed that only a small subset of them is significant to likelihood computation at any given time [7]. The hierarchical structure takes advantage of the sparseness by surveying the Gaussian pool in multiple resolutions given some acoustic feature vector \mathbf{x} . As part of the training process, the complete set of available Gaussian densities is clustered into a search tree, in which the leaves correspond to the individual Gaussians, and a parent node is the centroid of its children for a defined distance metric. At the bottom of this tree resides a many-to-one mapping, collapsing the individual Gaussians to the appropriate HMM state. Therefore, the HMM state s conditional likelihood of a given observation vector \mathbf{x} at time t is computed as

$$p(\mathbf{x}|s) = \sum_{g \in \mathcal{G}(s)} p(g|s) p(\mathbf{x}|g),$$

where $\mathcal{G}(s)$ is the set of Gaussians that make up the GMM for state s . Traversing the tree will yield a subset of active Gaussians, denoted by \mathcal{Y} . Based on \mathcal{Y} and the many-to-one mapping, the conditional likelihood of a state is approximated as

$$p(\mathbf{x}|s) = \max_{g \in \mathcal{Y} \cap \mathcal{G}(s)} p(g|s) p(\mathbf{x}|g). \quad (7)$$

If no Gaussian from a state is present in \mathcal{Y} , a default floor likelihood is assigned to that state.

To define the acoustic phonetic space used for speech activity detection, we apply an additional many-to-one mapping to the pruned result of the hierarchical tree. This many-to-one mapping is based on grouping sets of phonemes into three broadly defined classes: (i) the pure silence phoneme, trained from non-speech; (ii) the disfluent phonemes, which are noise-like phonemes, namely the unvoiced fricatives and plosives, i.e., the ARPAbet subset $\{b/, /d/, /g/, /k/, /p/, /t/, /l/, /s/, /sh/\}$; and (iii) all the remaining phonemes, such as the vowels and voiced fricatives. The three classes will be denoted by Sp_1 , Sp_2 , Sp_3 . From the acoustic feature \mathbf{x} , used to traverse the acoustic model hierarchy, we can form the speech detection class posteriors given in (8), for the three speech detection classes, as

$$Pr(Sp_i|\mathbf{x}) = \frac{1}{acc_mass} \sum_{g \in \mathcal{Y} \cap \mathcal{G}(Sp_i)} p(\mathbf{x}|g) p(g|Sp_i), \quad (8)$$

where

$$acc_mass = \sum_{i=1}^3 \left\{ \sum_{g \in \mathcal{Y} \cap \mathcal{G}(Sp_i)} p(\mathbf{x}|g) p(g|Sp_i) \right\},$$

and $\mathcal{G}(Sp_i)$ is the set of Gaussians defined by the mapping from phoneme to speech detection class Sp_i .

The process is illustrated in Fig. 2. Notice that the pruning at each level is accomplished using a threshold relative to the maximum scoring likelihood for that level [7]. As a result, the sharper the drop-off in Gaussian likelihoods, the more aggressive the pruning becomes. Therefore, both SNR and the phoneme being pronounced impacts the pruning. Features extracted from vowels and other voiced phonemes will result in more aggressive pruning than unvoiced fricatives, plosives and silence phonemes. This pruning will remain relative to SNR, with increasing SNR resulting in an overall more aggressive pruning.

The above observation results in additional speech detection features, based on class-normalized Gaussian counts. Denoting by N_{Sp_i} the number of Gaussians after hierarchical pruning that map to speech detection class Sp_i (see also Fig. 2), we

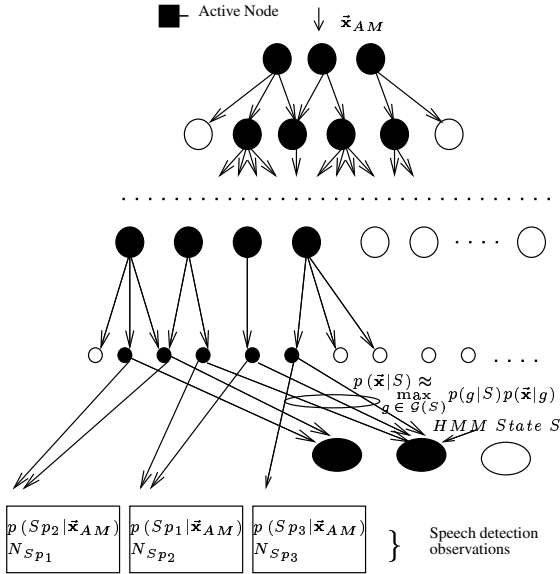


Figure 2: Hierarchical acoustic model and the corresponding acoustic phonetic speech detection observations

consider the normalized counts

$$\bar{N}_{Sp_i} = N_{Sp_i} / \sum_{j=1}^3 N_{Sp_j}, \quad \text{for } i = 1, 2, 3, \quad (9)$$

as additional features. Combining (8) and (9) we obtain the six-dimensional acoustic phonetic feature space at frame t given by $v_a(t)$, as defined in (10):

$$\begin{aligned} v_{a_i}(t) &= [\log(Pr(Sp_i|\mathbf{x})) \log(\bar{N}_{Sp_i})] \\ v_a(t) &= [v_{a_1}(t) \ v_{a_2}(t) \ v_{a_3}(t)]. \end{aligned} \quad (10)$$

2.3. Fusion of Acoustic Phonetic and Energy Features

From (6) and (10) we derive the fused feature space

$$v_f(t) = [v_e(t) \ v_a(t)]. \quad (11)$$

To this 11-dimensional feature space we apply principal component analysis (PCA), in order to decorrelate the features and to allow classification using GMMs with a diagonal Gaussian probability density, as discussed next.

2.4. Training and Classification

In the PCA step applied to (11), we choose as subspace the basis set formed by the eigenvectors corresponding to the top eight eigenvalues. This results to eight-dimensional features $v_p(t) = \mathbf{A} v_f(t)$, where \mathbf{A} denotes the PCA matrix. We subsequently train a three-class GMM classifier on vectors $v_p(t)$, i.e., for each speech detection class described in Section 2.2. This is accomplished by means of the expectation-maximization algorithm based on the training vector class labels, that are obtained by Viterbi alignment using the acoustic model, and the phoneme-to-speech detection class mapping (see also Fig. 3). In this work, eight mixture components are used for each class.

For classification, at each frame, we compute the GMM scores for the three classes, smoothed over a small number (ten) of consecutive frames, after extracting the necessary features, as described in Sections 2.1-3. In the final step, whenever a segment is classified as class Sp_2 , it is mapped to speech (Sp_3)

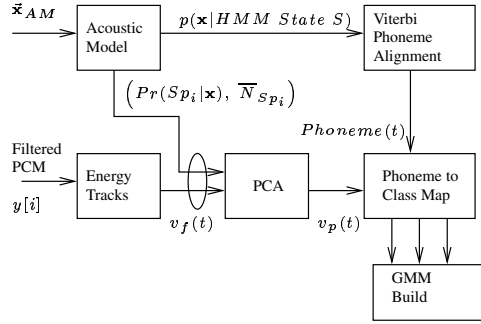


Figure 3: Training the three-class speech detection classifier

only if it lies between segments Sp_1 (silence) to its left and Sp_3 (speech) to its right, or vice-versa; otherwise it is mapped to silence (Sp_1). This is intended to handle both consonant-vowel-consonant transitions and non-stationary noise.

3. Experimental Results

Experimental results are presented for speech detection classifiers based on two acoustic models: (i) A narrowband telephony task, and (ii) a wideband far-field speech detection task.

3.1. Telephony

The narrowband telephony task consists of 8kHz audio from various landline, cellphone, and far-field speaker-phone devices. The acoustic model is trained on approximately 2000 hours of data from 20k speakers, using 60-dimensional acoustic features that result from 13-dimensional MFCCs followed by the application of temporal LDA and MLLT, and it consists of about 2.2k HMM states and 226k Gaussian mixtures. In this section we report experiments on two test sets of this task: The first consists of 8k utterances collected in high-energy, non-stationary noise environments, such as automobiles; the second contains 33k utterances recorded using handsets in a clean telephony environment.

In the first experiment, we compare three speech detection systems in the noisy test set. The first system is the one proposed in this paper, and trained as in Section 2.4. This is compared against two baseline approaches, one using solely energy based features [5], and the other using the acoustic-phonetic features of Section 2.2, in both cases, directly fed into a three-class GMM classifier. The experimental results are depicted in Fig. 4. There, we plot a receiver operating curve (ROC) depicting speech detection false acceptance (FA) versus false rejection (FR) rate, defined where errors exceed 600 ms and 300 ms respectively, as well as the total error, defined as the (FA+FR) minimum. As it becomes clear, the proposed method significantly outperforms the two baselines: For example, we observe a relative 12.4% gain in total error from 7.49% to 6.56%, when adding energy features to the acoustic phonetic ones. Fig. 4 also shows the energy only speech detector total error of 28.6%, (FA=0.28, FR=0.06). We only show this energy point, as this is the optimal setting for ASR performance on the 33k sentence clean telephony test set.

In the second experiment, shown in Table 1, we compare the fused speech detection system proposed in this paper to the energy-based only system [5] on the clean telephony test set. In addition to the FA and FR speech detection rates, ASR performance is also depicted, measured in word and sentence error rates (WER/SER), %. This is of interest, as typically a speech

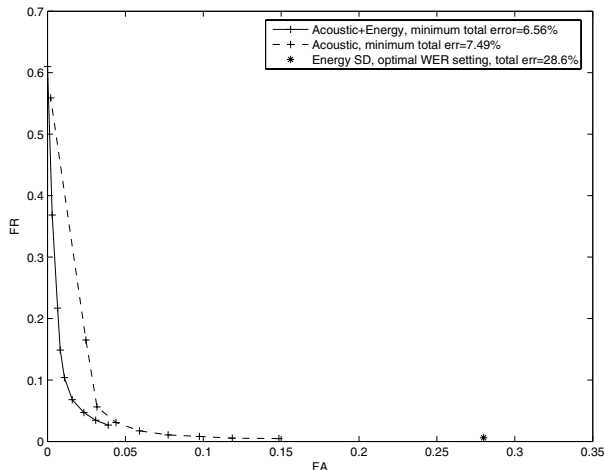


Figure 4: ROC comparing the proposed fused system with its purely energy or acoustic phonetic based subsystems for speech detection in the noisy telephony task. A single operating point that corresponds to optimal ASR performance is shown for the energy based speech detector

detector is used in conjunction with an ASR system. The energy based system is reported for two operating points, (a) corresponding to optimal ASR performance, and (b) corresponding to the minimum FA+FR point, as defined on the noisy test set. By comparing points (b) on the table, we clearly observe the superiority of the proposed algorithm, as measured by FA+FR, WER, and SER.

3.2. CHIL

The second evaluation task considered in this work is a wide-band, single-channel, far-field speech detection task for a seminar/lecture scenario. The data for this task are part of a larger seminar corpus, collected by the University of Karlsruhe, Germany, as part of the CHIL integrated project (“Computers in the Human Interaction Loop”), supported by the European Commission. The data have been recorded in a smart room, equipped with multiple acoustic and visual sensors, for example microphone arrays, close-talking microphones, fixed and steerable cameras [6]. The speech detection task has been evaluated in a subset of the CHIL seminar corpus by six CHIL consortium partners, based on the first audio channel of a linear, 16-channel microphone array, recorded at 16 kHz. The data consist of seven seminars, each by a different speaker, and contain mostly speech by the presenter, but also background noise and audience questions. The development and evaluation data for the task consist of approximately 10 mins each per seminar, totaling 1:08 and 1:09 hrs in duration, respectively.

In our experiments, we have benchmarked two speech activity detection systems on the CHIL evaluation set. Both follow the algorithm presented in Section 2, and are based on an acoustic model consisting of about 2k HMM states and 43k Gaussian mixtures, trained on approximately 400 hours of data from 1000 speakers. The acoustic model uses 40-dimensional features, resulting from 24-dimensional MFCCs, followed by the application of LDA and MLLT. A three-class GMM for speech activity detection is also trained on these data, as discussed in Section 2.4. The main difference of the two systems is that the first employs the original acoustic model and GMM classifier, whereas the second system uses supervised

System	FA	FR	WER	SER
Energy-based (a)	2.53	0.20	3.20	9.97
Energy-based (b)	0.07	15.24	8.90	16.84
Fused (b)	0.19	1.34	3.41	10.45

Table 1: Speech detection FA/FR, and corresponding ASR WER/SER (all reported %) compared for two operating points of the energy-based detector versus the proposed fused system, all reported on the clean telephony test set

maximum-a-posteriori adaptation of the acoustic model on the CHIL development set, before extracting features (10), and trains the GMM solely on CHIL data. Not surprisingly, the performance of the second system is significantly better than the first, reaching (SDER, NDER, ADER) = (10.01%, 11.92%, 10.96%), versus (16.70%, 16.43%, 16.57%) of the unadapted system. The reported metrics stand for speech, non-speech, and average detection error rates, the latter being ADER = (SDER+NDER)/2, all given at a “balanced” point that satisfies $|\text{SDER}-\text{NDER}| / (\text{SDER}+\text{NDER}) \leq 0.1$.

The reported performance of the second system on the CHIL task was superior to all five other systems evaluated by the CHIL consortium, achieving a 4% to 36% relative ADER reduction. All such systems used energy-based speech detection and/or acoustic features (such as MFCCs) for speech/silence classification. The reported results demonstrate the superiority of the acoustic phonetic based approach proposed in this paper.

4. Summary

In this paper we presented a novel approach to speech activity detection that augments energy based features with acoustic phonetic ones. In contrast to traditional systems that often use the frequency based speech representation to directly provide features to a speech/silence classifier, the proposed technique utilizes an acoustic model to provide likelihood based features to the detector using a phoneme grouping into three clusters. The algorithm is applied on two speech activity detection problems, where it is shown to outperform traditional approaches.

5. References

- [1] Li, Q., Zheng, J., Zhou, Q., and Lee, C.-H., “A robust, real-time endpoint detector with energy normalization for ASR in adverse environments,” *Proc. ICASSP*, pp. 233–236, 2001.
- [2] Martin, A., Charlet, D., and Mauuary, L., “Robust speech/non-speech detection using LDA applied to MFCC,” *Proc. ICASSP*, pp. 237–240, 2001.
- [3] Bou-Ghazale, S. and Assaleh, K., “A robust endpoint detection of speech for noisy environments with application to automatic speech recognition,” *Proc. ICASSP*, pp. 3808–3811, 2002.
- [4] Padrell, J., Macho, D., and Nadeu, C., “Robust speech activity detection using LDA applied to FF parameters,” *Proc. ICASSP*, vol. 1, pp. 557–560, 2005.
- [5] Monkowski, M., *Automatic Gain Control in a Speech Recognition System*, U.S. Patent US6314396.
- [6] Macho, D., Padrell, J., Abad, A., et al., “Automatic speech activity detection, source localization, and speech recognition on the CHIL seminar corpus,” *Proc. ICME*, 2005.
- [7] Novak, M., Gopinath, R.A., and Sedivy, J., “Efficient hierarchical labeler algorithm for Gaussian likelihoods computation in resource constrained speech recognition systems,” available on-line at: <http://www.research.ibm.com/people/r/rameshg/novak-icassp2002.ps>