LOGISTIC DISCRIMINATIVE SPEECH DETECTORS USING POSTERIOR SNR

Arun C. Surendran[1] *Somsak Sukittanon*[2] *John Platt*[1]

[1]Microsoft Research, Redmond WA 98052, {acsuren, jplatt}@microsoft.com
[2]University of Washington, Seattle WA 98195, ssukitta@ee.washington.edu

ABSTRACT

We introduce an elegant and novel design for a speech detector which estimates the probability of the presence of speech in each time-frequency bin, as well as in each frame. The proposed system uses discriminative estimators based on logistic regression, and incorporates spectral and temporal correlations in the same framework. The detector is flexible enough to be configured in a single level or a "stacked" bi-level architecture depending on the needs of the application. An important part of the proposed design is the use of a new set of features: the normalized logarithm of the estimated posterior signal-to-noise ratio. These can be easily and automatically generated by tracking the noise spectrum online. We present results on the AURORA database to demonstrate that the overall design is simple, flexible and effective.

1. INTRODUCTION

Detecting the presence of speech is crucial in many applications. In some, like noise adaptation and speech enhancement, a simple presence (or absence) decision alone does not suffice - it is critical to estimate the *probability of the presence of speech* in each time-frequency bin as well as each frame [2]. In other applications, a simple frame level decision is sufficient, but the requirements may vary e.g. in source localization the probability of false detection (classification of noise-only frames as speech frames) should be low, whereas in speech coding a high speech detection rate is desirable. Thus an ideal system is one that produces *calibrated* probabilities i.e. measures that accurately reflect the actual frequency of occurance of the event (presence of speech). Such a system can (1) make decisions optimally based on utility theory, and (2) combine decisions from independent sources using very simple rules. Further, an ideal system should also be simple and be light on the use of resources.

In this paper, we introduce an elegant and effective design for a detector which can accurately estimate calibrated probabilities of the presence of speech based on logistic regression based classifiers. The design of our system is flexible enough to allow the arrangement of detectors in a cascaded or a uni-level architecture depending on the need of the application, without sacrificing performance. The cascaded version first detects the probability of presence of speech in each time-frequency atom, and uses these values to make a frame level probability estimate. The uncascaded version estimates these probabilities directly from the feature data. One important aspect of the design is the use of features based on posterior signal-to-noise ratio. These features are designed to fit well with the detector, and can be easily generated on-line. The features and the classifier put together, make a simple yet effective speech detector.

The paper is organized as follows: We describe the classifier/detector architecture in Section 2. We describe the proposed feature in Section 3. Finally, in Section 4, we present results on the AURORA database that demonstrate the effectiveness of the proposed technique.

2. LOGISTIC REGRESSORS FOR SPEECH DETECTION

2.1. Previous Work

Many approaches have been proposed to detect speech presence or estimate its probability at the frame level. One very popular method is to use likelihood ratio (LR) tests based on Gaussian models. A voice activity detector using such a test was proposed in [5]. In essence, it uses a smoothed signal to noise (SNR) ratio estimate of each frame to implement this test, and seems to be effective for speech detection. Unfortunately, it (like other LR based tests) suffers from the problem of threshold selection, and the LR scores do not translate easily into true class probabilities. To convert from the former to the latter, additional information in the form of prior probabilities of the hypotheses need to be known. Further, this method assumes that both noise and speech have normal distributions with zero mean, which seem to be overly restrictive assumptions. In the rest of this paper we refer to this method as the "Gaussian" approach, and compare our technique with it. LR tests can be improved with larger mixture models, but these are computationally expensive.

Other techniques make speech / non-speech decisions at the frame level (i.e. they estimate a 0/1 indicator function), and smooth this over time to estimate the probabilities [2]. Some others use hard or soft voting mechanisms on top of such indicator functions estimated at the time-frequency atom level [3]. One technique that is frequently used to estimate probabilities is a linear estimation model: $p = A + B\mathcal{X}$, where p is the probability, $\mathcal{X}$ is the input (this could be LR scores or observed features like energies), and A and B are the parameters to be estimated. One of the probability estimators in [2], even though not explicitly formulated this way, effectively adopts the linear model and uses the log of smoothed energy as the input. The two most important problems with the linear model are that the predicted probabilities can be greater than 1 or less than 0 (high and low thresholds must be set right to avoid this problem), and the variance of the error in estimation depends on the input variable.

2.2. Logistic Regression

As an alternative to models discussed above, we propose the use of a "logit" or a logistic regression model [6] for speech detection.

The class probability is estimated as:

$$p_{\mathcal{X}} = \frac{1}{1 + \exp(-A - B\mathcal{X})}, \quad (1)$$

where $\mathcal{X}$ is the input and A and B are the parameters of the system. These parameters are estimated by minimizing the cross-entropy error function [6]:

$$\mathcal{E} = -\sum_{\mathcal{X}} t_{\mathcal{X}} \log(p_{\mathcal{X}}) + (1 - t_{\mathcal{X}}) \log(1 - p_{\mathcal{X}}), \quad (2)$$

where $t_{\mathcal{X}}$s are the target labels for the training data $\mathcal{X}$, and hence is discriminative. This also provides the maximum likelihood estimate of the class probability.

The logit function provides very good estimates of the posterior probability of the membership of a class $p(\mathcal{C}|\mathcal{X})$ for a wide variety of class conditional densities of the data $\mathcal{X}$ [6]. If the densities are multivariate Gaussians with equal variances, then this estimate is the exact posterior probability. But to emphasize, normality is not a necessary condition for logit to be effective. Further, the logit function does not run into all the other problems associated with the Gaussian models mentioned in the previous paragraph. The model has many other advantages: The parameters can be easily calculated using gradient descent based learning algorithms. If the input vector $\mathcal{X}$ contains data from adjacent time and frequency atoms, the logit function becomes an easy way to incorporate both temporal and spectral correlation into the decision without overly worrying about the underlying distributions. To summarize, the attraction of this technique is its richness and simplicity.

2.3. Stacked Architecture

One additional advantage of the logit model is the flexibility it provides in engineering a useful architecture. Some applications need the probability of speech to be estimated at both the time-frequency atom level and at the frame level. In a classification problem, it is known that the transform of a variable that has minimum Bayes risk is the one that estimates the posterior class probability of its true class [8]. Looking at the frame level detection from this perspective, we can easily see that the conditional class probabilities for each time-frequency atom is the best possible input for the frame level detector. This suggests a bi-level architecture where the first level of detectors operate at the atom level, and the outputs of these are used as inputs to a frame level speech detector. Posterior class probabilities have been used as features for HMM based speech recognition (e.g. [9]).

The first level has one detector for each atom. The input to each of these detectors is the vector $\mathcal{X}(k,t) = [\mathcal{Y}(k-a : k+a, t-i : t+i)]$ which is the concatenation of the feature $\mathcal{Y}$ in the time-frequency neighbourhood of the relevant time-frequency bin (k,t). If a delay in processing cannot be tolerated, the feature can be strictly causal and need not include future frames. For the single-layer system, $\mathcal{X}$ in the above equation is substituted by its mel-band derivative (see Section 3.1).

Once the first layer of detectors compute the probabilties $\mathcal{P}(k,t)$, these can be concatenated in a fashion similar to $\mathcal{X}$ and fed to a single detector in the second layer. This architecture has some similarities to the "stacked generalizer" proposed by Wolpert [7] for minimizing bias in learning. In Wolpert's work, the first and second layers both estimate the same function, except they operate on different input spaces. Thus our proposed classifier can be called "stacked" if we interpret each atom level detector to be a "poor" predictor of the frame level speech presence.

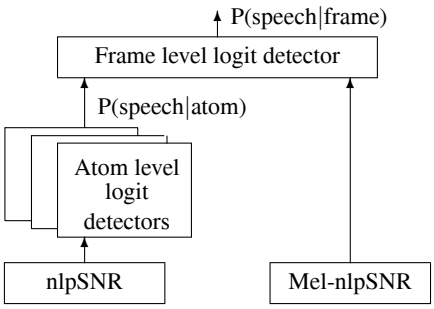

Fig. 1. Stacked (left) or an unstacked (right) architectures can be used based on the needs of the application.

3. POSTERIOR SNR BASED FEATURES

An important desired characteristic of a speech detector is that the features it uses remain sufficiently simple. Complex features derived solely for the purpose of detection add to the computational complexity. In this paper, we introduce a very simple set of features based on the estimated signal-to-noise ratio (SNR). Since the actual SNR of each frame can be known only by estimating the actual speech and noise components of each given noisy frame, it is easier to deal with the *estimated posterior SNR* which is the ratio of the energy in the given frame Y to *estimated* noise energy $\hat{\lambda}$: $\xi(k,t) = |Y(k,t)|^2/\hat{\lambda}(k,t)$, where k,t are the frequency and time indices respectively. The terminology used here was first proposed by McAulay and Malpass [1]. The estimate of the actual SNR (also referred to as the *prior SNR* [1]) can also be used as a feature. The likelihood ratio based method proposed in [5], in fact, uses an estimate of such a feature. Our preliminary experiments show that this feature performs better than the posterior SNR based feature. But estimating this is equivalent to doing speech enhancement [2, 3], and can be complicated. Hence, for the sake of simplicity, we will not use it in this paper. To compute the spectrum we use modulated complex lapped transforms (MCLTs) [10]. MCLT is a particular form of cosine modulated filter-bank that allows for perfect reconstruction. FFTs can easily be used instead of MCLTs without changing any other procedure in this paper.

Since the features are being fed to a learning machine, some preprocessing is needed to improve generalization and learning accuracy. First, since short term spectra of speech are modeled well by log-normal distributions, we use the logarithm of the SNR estimate, rather than the SNR estimate itself. Second, we normalize the input so that its variance is 1. In this paper, we precompute the variance for each coefficient over the training set and use it as the normalizing factor. Thus our new feature is a normalized logarithm of the estimated posterior SNR (nlpSNR), and is written as:

$$\mathcal{Y}(k,t) = \frac{1}{2\sigma(k)}\{\log|Y(k,t)|^2 - \log\hat{\lambda}(k,t)\},$$

where $\sigma(k)$ is a normalizing factor, and all other variables are as defined earlier in this section.

3.1. Mel-transform for single layer system

For a single layer system which does not need atom level decisions, we can use features that have poorer frequency resolution. For

example, both $|Y|^2$ and $\hat{\lambda}$ can be converted into Mel-band energies before nlpSNR is computed.

3.2. Automatic noise tracking

The noise power λ can be tracked using various algorithms [2, 4]. In this work, we use a two level online automatic noise tracker. The initial estimate is bootstrapped using a minima tracker (e.g. [4]) which is used to compute the probability of the presence of speech similar to the method in [2]. Following that, a maximum *a posteriori* estimate of the noise spectrum is obtained. Since the focus of this paper is not the estimate of the noise, we will eschew discussing it in further detail.

4. EXPERIMENTS AND RESULTS

4.1. Database

We use the well known AURORA database [11] for our experiments. The database has spoken digits from male and female speakers with different types of noises added to clean signals at 20, 15, 10 and 5dB SNR levels. We are interested only in two subsets of the database named TESTA and TESTB. The type of noise in TESTA and TESTB are different, though the speakers are the same. We chose 100 male and 100 female speaker data from TESTA for training the logistic regressors; 10 male and 10 female speaker data from TESTA for validation during training, and 100 male and 100 female speaker data from TESTB for testing. We ensure that the speakers selected for testing from TESTB are entirely different from those in the training set. Also the data at 5dB SNR is not used for training. In total, about 55000 frames are available for training and over 68000 frames for testing.

The knowledge that this is a "stereo" database i.e. the data contains "clean" signals and their corresponding noisy counterparts, is used *only* to generate the true labels at the atom and frame level needed for training. This information was *not* used for online noise estimation or in testing.

The true labels were generated by thresholding the energy in each time-frequency bin/frame of the clean data. The thresholds were selected so that all speech events were retained. This was verified through listening experiments on a small fraction of the training data. The threshold was tuned so that the low energy speech events and the transitions just barely made the cut.

A 128-pt MCLT was used to compute the spectrum every 16ms using a 32ms window. For each time index t, the input vector to the logit functions contained all the spectral components of the feature vector at t and its immediate neighbors in time ($t-1$ and $t+1$). All the logit parameters in this paper were estimated using stochastic gradient learning [6].

4.2. Comparing Features

The first set of experiments demonstrate the effectiveness of the nlpSNR feature over other spectral features. The classifiers were single layer logistic regressors for frame level speech detection. The features are transformed into 20 mel-band energies for classification. We choose two other features for comparison, each with a different method for spectral normalization: (1) variance normalized noisy speech spectra (referred to as "Y (normalized)", and (2) noise normalized spectra (referred to as "Y (min-max)", based on maxima and minima tracking in the spirit of [4] except that feature is normalized thus: $\mathcal{Y} = (Y(t) - Y_{min})/(Y_{max} - Y_{min})$. Y_{max} and Y_{min} are calculated over a recent time period so that they can track the local signal and noise variations. The results are shown in Figure 2. The plots show the ROC curve (correct detection of speech frames vs. false alarm). The "minimum error" numbers, where the weighted error (false alarm*(ratio of noise frames)+ false rejetion*(ratio of speech frames)) is the least, are also listed for each graph. The axes are shortened to highlight the upper-left quadrant of the plane. We can clearly see that the new feature outperforms the others at all SNRs.

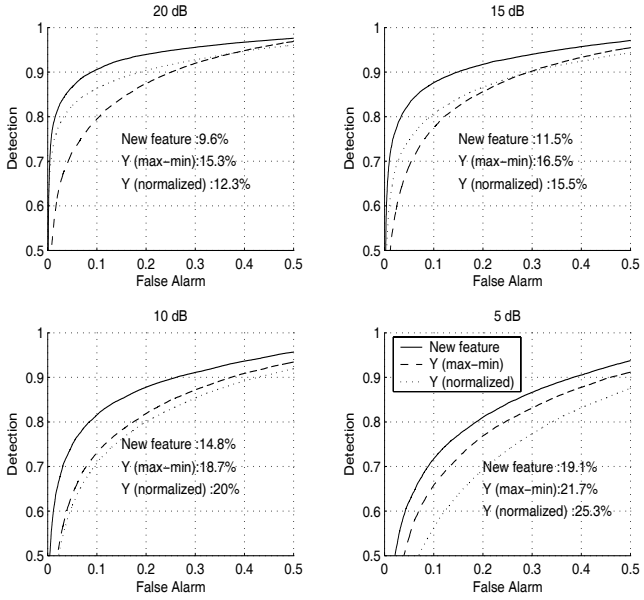

Fig. 2. ROC curves comparing features at various SNRs: (1) New nlpSNR feature (solid line), (2) Signal power with elementary noise compensation (dashed line), and (3) Variance normalized signal power (dotted line). Minimum error numbers are inscribed.

4.3. Atom level decisions

Now we demonstrate the strength of the classifier. We compare our proposed approach to the "Gaussian" method mentioned earlier [5]. The first set of experiments are at the atom level. We use 128 detectors - one for each bin. Here we only show results from one frequency bin - the one centered at 1000 Hz. The results at this bin summarize the performance in most other bins. The rationale for this statement will be clear as we explain the results. ROC curves for four different SNRs are shown in Figure 3. The SNR labels in the figure do not reflect the SNR in the bin, but the SNR of the entire file. So for frequency bins that are not covered by the noise spectrum, the bin SNR will be much higher. We can see from the figure that in bins with high SNR (e.g. 20dB) the Gaussian method is slightly better. But as the SNR worsens, the proposed method significantly outperforms it. In general, in this database, the bin SNRs are higher at very high indices. So for frequency bins closer to 4KHz, the Gaussian does slightly better e.g. at 3KHz, the average minimum error over 20dB-5dB data range is 9.85% for the new system versus 8.35% for the Gaussian. At the same time, in frequency bins including a lot of noise, the new method is much

better e.g. in the 500Hz bin the new method has a minimum error of 15.38% vs. 23.35% for the Gaussian.

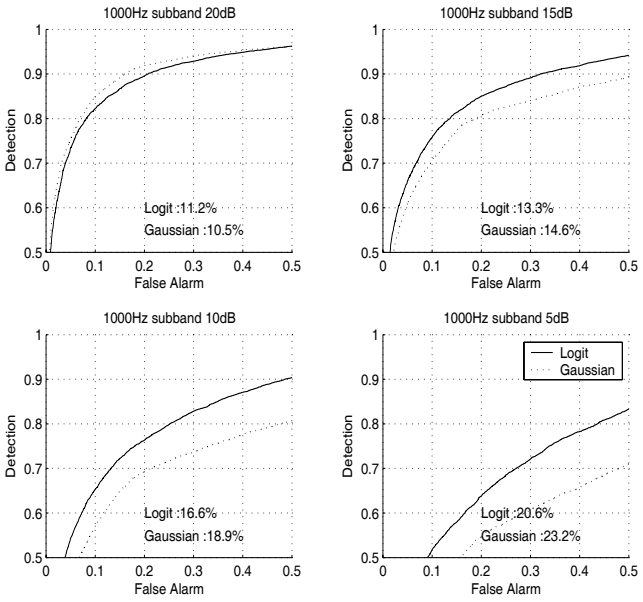

Fig. 3. ROC curves for detection of speech in the frequency band around 1000Hz at various SNRs using (1) the proposed logit detector (solid line), and (2) the Gaussian method [5] (dotted line). Minimum error numbers are inscribed.

4.4. Frame level decisions, Stacked vs. Single level

In this section, we compare classifiers at the frame level. Figure 4 shows the ROC curves for three cases: the new method using both a stacked and an un-stacked architecture, and for the Gaussian method. The stacked and the unstacked methods perform very similarly and both of them outperform the Gaussian method in all parts of the ROC curve. Hence the new method will be better than the Gaussian method for any application.

Since the performance of the stacked and single layer system are so close, the choice of architecture can be made based only on the application. If atom level decisions are also needed in addition to frame level decisions, then the stacked classifier can be used; otherwise the single level network will suffice.

One advantage of a logistic regression system is that it is possible to evaluate the influence of each component of the input vector by analyzing the significance of the corresponding weight (using for example the Wald statistic). Through such an analysis, it is possible to use fewer input coefficents. Due to lack of space we postpone a detailed discussion for future publications.

5. SUMMARY

We present a simple, yet effective solution for estimating the probabilty of the presence of speech at the frame level using logistic regression based detectors. We propose a new feature (nlpSNR) which is easy to estimate online and aids detection significantly. It is also designed to fit well with our back-end classifier. The detector is flexible, so we can choose to implement it either in a stacked architecture or in a uni-level architecture based on the needs of the application. Both these methods are equally effective. We present convincing results on the AURORA database to demonstrate the strength and flexibility of the new approach.

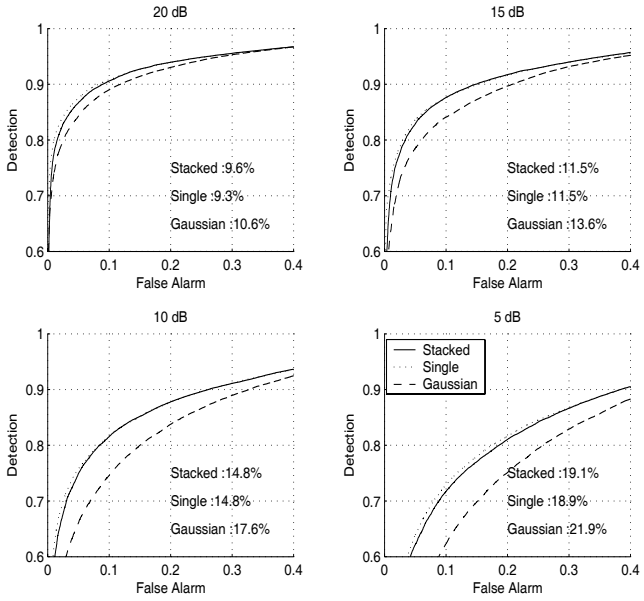

Fig. 4. ROC curves comparing the new design at various SNRs using (1) a stacked architecture (solid line), (2) a single layer architecture (dotted line) and (3) the Gaussian based approach [5] (dashed line). Minimum error numbers are inscribed.

6. REFERENCES

[1] R. J. McAulay et. al., "Speech enhancement using a soft-decision noise suppression filter", ASSP-28, pp. 137-145, Apr '80.

[2] I. Cohen et. al., "Speech enhancement for non-stationary noise environments", Sig. Proc. 81 (2001) pp. 2403-2418.

[3] D. Malah, et. al., "Tracking speech presence uncertainty to improve speech enhancement in noisy env.", ICASSP '03.

[4] R. Martin, "Spectral subtraction based on minimum statistics," Proc. EUSIPCO-94, pp. 1182-1185, Edinburgh, 1994.

[5] J. Sohn et. al., "Statistical model based voice activity detector", Sig. Proc. Letters, V.6, No.1, pp. 1-3, Jan '99.

[6] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press.

[7] D. H. Wolpert, "Stacked generalization", Neural Networks, Vol. 5, pp. 241-259, Pergamon Press, 1992.

[8] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd Edition, Springer-Verlag.

[9] H. Hermansky, et. al., "Tandem Connectionist Feature Extraction for HMM Systems", ICASSP 2000, Istanbul.

[10] H. S. Malvar, "A modulated complex lapped transform and its applications to audio proc.", ICASSP '99, pp. 1421-1424.

[11] H. G. Hirsch, et. al., "The AURORA Experimental Framework", ASR2000, Paris, France, Sep 2000.