A Robust Multiple Feature Approach To Endpoint Detection In Car Environment Based On Advanced Classifiers

C. Comas[1], E. Monte-Moreno[1], J. Solé-Casals[2]

[1]TALP Research Center
Universitat Politècnica de Catalunya (Spain)
enric@gps.tsc.upc.es
[2]Signal Processing Group, University of Vic (Spain)
jordi.sole@uvic.es

Abstract. In this paper we propose an endpoint detection system based on the use of several features extracted from each speech frame, followed by a robust classifier (i.e Adaboost and Bagging of decision trees, and a multilayer perceptron) and a finite state automata (FSA). We present results for four different classifiers. The FSA module consisted of a 4-state decision logic that filtered false alarms and false positives. We compare the use of four different classifiers in this task. The look ahead of the method that we propose was of 7 frames, which are the number of frames that maximized the accuracy of the system. The system was tested with real signals recorded inside a car, with signal to noise ratio that ranged from 6 dB to 30dB. Finally we present experimental results demonstrating that the system yields robust endpoint detection.

1. Introduction

In speech and speaker recognition a fast and accurate detection of the speech signal in noise environment is important because the presence of non-voice segments or the omission of voice segments can degrade the recognition performance [2,3,4]. On the other hand in a noise environment there are a set of phonemes that are easily masked, and the problem of detecting the presence of voice cannot be solved easily. The problem is further complicated by the fact that the noise in the environment can be time variant and can have different spectral properties and energy variations. Also there are limitations on the admissible delay between the input signal, and the decision of the presence or absence of voice. Therefore the variability of the environment justifies the use of different features, which might be adapted to discriminate voice from different kind of environmental noise sources. The variability of circumstances justifies the use of the aggregation of classifiers which are trained differently. The use of

This work has been partially supported by the Spanish CICyT project ALIADO, the EU integrated project CHIL and the University of Vic under the grant R0912.

FSA is justified by the fact that the classifiers make bursts of nearly consecutive mistakes; these bursts of false alarms or false positive decision are easily filtered by the FSA.

2. General structure of the system

We designed a system which consisted of a frame level classifier followed by a FSA. The idea behind the design was to use a robust classifier with input features, which in isolation have proved to yield good performance in different on the environmental conditions, and a FSA which implemented the decision logic that filters short bursts of nearly consecutive false alarms or false positives.

2.1. Selected Features

The selected features were:

- The Teager energy [1] is a measure of the energy of a system in a simple harmonic motion, which is $E \propto A^2\omega^2$. In [2] this measure is proposed for endpoint detection. We used the windowed mean of this energy of frame 'j':

$$TE_j = \sum_{i=1}^{n}\left[x^2(i) - x(i+1)x(i-1)\right] w(i) \quad \textbf{(1)}$$

- Differential Teager energy: The derivative of the Teager energy was computed by filtering the Teager energy with: $H(z) = 1 - z^{-2}$
- Zero crossing rate: was computed for each frame by

$$CX_j = \sum_{i=1}^{n}\left|\ \mathrm{sign}(x(i)) - \mathrm{sign}(x(i-1))\ \right| w(i) \quad \textbf{(2)}$$

- Spectral entropy: [3]: This method is based on measuring the normalized spectral power density function for each frame, which can be interpreted as the probability of a certain frequency. The associated entropy is computed, and non discriminative frequencies are windowed out

$$H_j = -\sum_{k=1}^{n} w_k\, p_k \log p_k \quad \textbf{(3)}$$

In case of speech, for certain phonemes, the energy is concentrated in a few frequency bands, and therefore will have low entropy, while in the case of noise with flat spectrum or low pass noise, the entropy will be higher.

- Spectral coherence between consecutive frames gives a measure of the similarity between of two frames

$$\left|\gamma(\omega)\right|^2 = \frac{S_{xy}(\omega)S_{yx}(\omega)}{S_x(\omega)S_y(\omega)} \quad \textbf{(4)}$$

We use as feature the values of $\left|\gamma(\omega)\right|^2$ computed by means of the DFT.

2.2. Classifiers

We did experiments with four different classifiers: a linear discriminant, the AdaBoost of linear classifiers the bagging of decision trees, and a Multilayer Perceptron (MLP). The linear discriminant was selected as benchmark. As can be seen in figure 1, when the classification (with AdaBoost) is done at the frame level without the finite state automata, there is a high number of false alarms. This phenomenon was common to all classifiers. The distribution of the false alarms is such that they can be easily filtered by the finite state automata. Nevertheless, it also means that at the frame level the confusion between classes is high. The use of a decision tree alone was discarded because of the poor results, when the tests were done with noisy signals. The decision trees in these cases grow specific branches for similar cases with different labels, which means that the decision trees tend to grow in excess. Pruning the trees degraded abruptly the performance. The best results with a decision tree were slightly worse than the linear discriminant.

We do not present results with support vector machines either, because the high overlap between classes in the feature space in the low SNR case, yields bad results and an extremely high number of support vectors.

The use of bagging of the decision trees is justified because of this low ‘hit’ rate at frame level. Bagging decision trees improved the performance because we were able to grow trees (trained by bootstrapping on the training database) with a high number of nodes, i.e. adapted to the specificities of the training database, and afterwards the aggregation of trees smoothed the variance and therefore reduced the error rate [5].

The AdaBoost [6] was selected in order to improve the accuracy of a linear discriminant. As the classification rate of a linear discriminant is low, the use of AdaBoost creates a set of classifiers that specialize in the distribution of the misclassified frames in the feature space. Also in the case of AdaBoost it is known that the performance degrades when there is a high overlap between cases in the feature space, consequently in order to reduce the degradation caused by this overlap, we selected the number of classifiers by cross validation with a criterion based on the final recognition rate, i.e. the accuracy after the finite state automata. The use of AdaBoost in combination with decision trees was discarded because the preliminary experiments yielded extremely big trees, which were computationally prohibitive. This is explained because the high overlap between classes made the trees to specialize in contradictory examples.

2.3. The FSA and the decision criterion

After the classifier step we used as decision logic a FSA, which had four states, and the transition from one state to the other was controlled by the number of frames that corresponded to each class: 1 →voice, 0 →no voice. A diagram is shown in figure 3. The parameters for the FSA: **N** = number of contiguous voice frames, **M** = number of contiguous non voice frames; were tuned by a compromise between: the number of consecutive frames of false negatives, frames of false positives, and the rate of real negatives (absence of voice) and real positives (presence of voice) after applying the FSA. The rates were computed on the training database. The compromise was obtained by inspection of the ROC curve (Receiver Operating Characteristics), (see figure 4) and histograms of duration of bursts of false negatives, and false positives. These parameters were adjusted as follows:

We plotted histograms of: a-consecutive false positive frames, and real positive (consecutive frames of voice), for different values of M, and b-histograms of consecutive false negative frames, as compared with the real negative frames (absence of voice), for different values of M.

We selected the value of M that in both cases corresponded approximately to the crossing of the false positive and real positive cases. Then with the value of M fixed, we plotted the ROC curve for different values of N and selected a value of N that gave a hit rate of 98%.

3. EXPERIMENTAL RESULTS

The experiments were done with a subset of the SpeechDat Car [7], in Spanish. The subset consisted of 100 files with speakers of both sexes, and phrases of different kinds. The files were hand labeled. Each file contained four recordings of the same signal, one recorded near the mouth of the speaker, which had a mean SNR of about 30dB. The other three were recorded in different places of the car, and had a mean SNR of about 8,5 dB. The sampling frequency was 16 kHz. The signal was divided into frames of 33 ms, with an overlap of 50%. A pre-emphasis was done before the processing of the signal. For each frame we computed the following features: Teager energy, Differential Teager energy, zero crossings, spectral entropy, and spectral coherence between successive frames. Then the classifier made a decision at frame level, and the FSA made a final labeling based on the history of the decisions made by the classifier. Therefore the delay introduced by the system is the delay of the FSA, which is of 7 frames for ‘in speech’ and ‘out speech’.

The endpoint detection results were computed by a 5 fold cross validation. This was done in order not to make the results dependent of the random distribution of files between training and testing. This gave a mean of 13500 for train and 3300 frames for test in each cycle of the cross validation process. Of these frames approximately 65% was no-voice and 35% voice.

The number of linear discriminants in the AdaBoost was selected to be 10 after experiments with cross validation on the training database. For comparison purposes, the number of trees in the bagging experiment was selected also to be 10. Increasing the number of bagged trees did not give significant improvements. The topology of the MLP was decided by the performance on the training database, and the training was stopped by the performance on a validation subset of the training database.

In order to quantify the performance of the system we will use the following parameters:

- Accuracy (ACC), the ratio between the total numbers of predictions that were correct to the total number of cases.
- True positive rate (TPR), which is the ratio of positive cases that were correctly classified to the total number of positive cases.
- False positive rate (FPR), which is the ratio of negative cases that were incorrectly classified as positive to the total number of negative cases.
- Precision (PRC), which is the ratio of predicted positive cases that were correct to total number of positive cases.

The results are presented in table I and II. We have not presented the confidence margins. Due to the high number of frames, in all cases they lower than 10e-3. The results have to be interpreted in the light of figure 5. Most of the errors came from a misplacement of the transition between voice/no voice, which in most cases can be quantified in a few frames. The manual labeling of the boundary between the segments of voice/ no voice is subjective in a margin of a few frames. This is reflected in the fact that the histogram of placement of the boundary is slightly biased to the right or left depending on the case of beginning or ending of speech. This bias comes from the fact that at this point the classifiers gives bursts of false alarms, which are filtered by the FSA. Figure 2 shows an example of this bias on the endpoints.

Table 1. Results (%) of a channel with mean SNR of 30dB

	Linear Classifier	AdaBoost linear classifiers	Bagging decision trees	MLP(5;3;1)
ACC	86	92	92	93
TPR	77	89	89	91
FPR	0.6	3	2	2
PRC	99	97	98	98

Table 2. Results (%) of a channel with mean SNR of 8dB

	Linear Classifier	AdaBoost linear classifiers	Bagging decision trees	MLP(5;3;1)
ACC.	78	83	81	84
TPR	77	85	82	86
FPR	19	18	20	18
PRC.	85	87	86	87

4. CONCLUSIONS

In this paper, we propose a system for real-time endpoint detection. The system is based in a frame level classifier followed by a finite state automaton that filters false alarms or false positives. The frame level classifier is based in five different features. The delay introduced is of 7 frames. The experiments were done on speech recorded in a car environment. In the future we will evaluate the endpoint detection with a speech recognition system.

References

1- J.F. Kaiser, "On a Simple Algorithm to Calculate the Energy of a Signal," *Proc. ICASSP*, 381-384, 1990.

2- G. S. Ying, C. D. Mitchell, and L. H. Jamieson, “Endpoint Detection of Isolated Utterances Based on a Modified Teager Energy Measurement”, *Proc. ICASSP,* II.732-735, 1993.

3- Jia-lin Shen, Jeih-weih Hung, Lin-shan Lee, "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments", *Proc. ICSLP* CD-ROM 1998.

4- W.-H.Shin, B.-S.Lee,Y.-K.Lee,J.-S.Lee, “Speech/Non-Speech Classification Using Multiple Features For Robust Endpoint Detection”, *Proc. ICASSP*, 1399-1402, 2000.

5- L. Breiman. “Bagging predictors”. *Machine Learning*, 24(2):123-140, 1996

6- Y. Freund and R. E. Schapire.“Experiments with a new boosting algorithm”. *Proc. 13 th International Conference*, pp 148--156. Morgan Kaufmann, 1996.

7- Asunción Moreno, Borge Lindberg, Christoph Draxler, Gaël Richard, Khalid Choukri, Stephan Euler, Jeff Allen “SPEECH DAT CAR. A Large Speech Database For Automotive Environments” *Proc. of the II Language Resources European Conference*. Athens, 2000

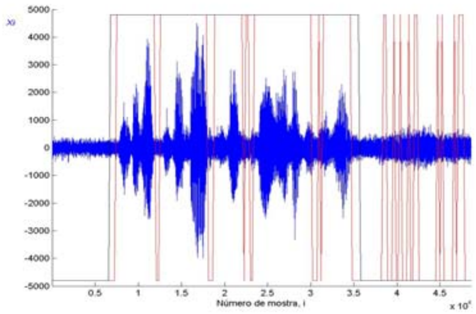

Fig. 1. Segmentation done by a frame level classifier (AdaBoost), for the utterance: ‘seleccionar centro de la ciudad’. Note the false alarms and false negative points.

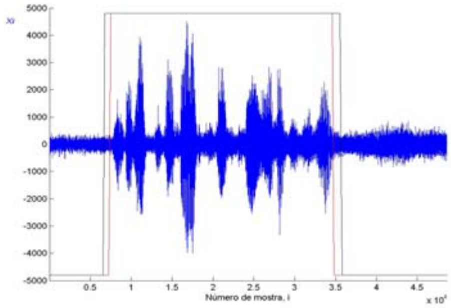

Fig. 2. Segmentation done by the FSA on the frame level classification. Comparison on an example of a manual segmentation (dark line) with the segmentation done by our system (clear line)

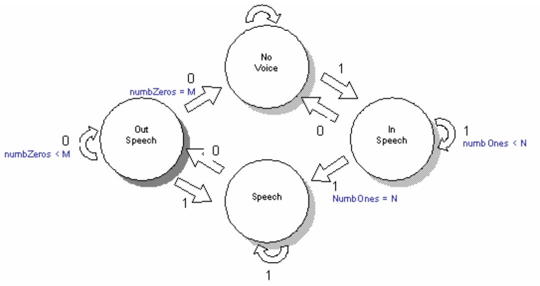

Fig. 3. Finite State Automata that filters the bursts of false alarms and false negatives

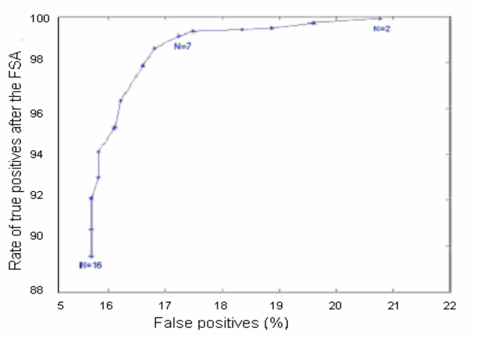

Fig. 4. The ROC obtained by the use of the FSA, with M=7 frames as out-speech for the channel with a SNR=8,5dB. The frame level classifier was based on the AdaBoost

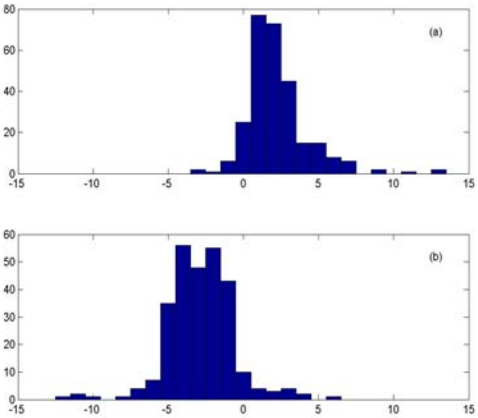

Fig. 5. Histogram of the distance of a frame to the nearest endpoint. Upper histogram is the transition from 'no voice' to voice. Lower from voice to 'no voice'