ROBUST SPEECH ACTIVITY DETECTION USING LDA APPLIED TO FF PARAMETERS

Jaume Padrell, Dušan Macho and Climent Nadeu

TALP Research Center
Universitat Politècnica de Catalunya
{jaume,dusan,climent}@talp.upc.es

ABSTRACT

Speech detection becomes more complicated when performed in noisy and reverberant environments like e.g. smart rooms. In this work, we design a robust speech activity detection (SAD) algorithm and we evaluate it on distant microphone signals acquired in a smart room-like environment. The algorithm is based on a measure obtained from applying Linear Discriminant Analysis on Frequency Filtering features. With a time sequence of this measure, a decision tree based speech/non-speech classifier is trained. The proposed SAD system is evaluated together with other SAD systems (GSM SAD and ETSI Advanced Front-End standard SAD) using a set of general SAD metrics as well as using the ASR accuracy as a metric. The proposed SAD algorithm shows better average results than the other tested SAD systems for both the set of general SAD metrics and the ASR performance.

1. INTRODUCTION

The purpose of Speech Activity Detection (SAD) is to detect speech in a continuous stream of audio signal. SAD is being used as a supporting technology in many speech related technologies like automatic speech recognition, speech coding, speaker identification, speaker localization, etc. SAD can save computational resources (and batteries) in the devices where the processing of non-speech events is not needed. In the case of ASR, besides saving the resources, SAD can significantly improve the recognition performance of the system if the non-speech events are excluded from the recognition process. On the other hand, in many speech enhancement technologies, the reliable detection of non-speech portions of signal is of interest in order to properly estimate the noise characteristics.

In our research center, we start to work with a smart room environment. In a smart room, the audio acquisition is assumed to be done in an unobtrusive way, usually by a network of far-field microphones. The distance of these microphones from the audio source (speaker) can vary from several centimeters to several meters. In such challenging environment, a high robustness of all speech technologies, including SAD, against environmental noises and reverberation is extremely important. In this work, we propose and test a SAD algorithm for this environment; it is based on robust speech features, Frequency Filtered log spectral energies further processed by Linear Discriminant Analysis. The speech activity measure that is obtained in this way is described in detail in Section 2. Section 3 describes the speech/non-speech decision taking block. In Section 4, the performance of our SAD and several other SAD systems is compared in terms of a) coincidence with the reference speech/non-speech labels and b) the ASR accuracy.

2. FF+LDA MEASURE

The presented SAD system is based on Frequency Filtering (FF) features. In [2], higher robustness of FF features in comparison to MFCC features was reported in noisy ASR tests; we expect this robustness aspect of FF to be reflected in our SAD system.

The FF feature extraction scheme consist in calculating a log filter-bank energy vector for each signal frame and then applying a FIR filter h(k) on this vector along the frequency axis. We used the h(k)={1, 0, -1} filter in our SAD. Notice that FF requires less computation than MFCC.

The initial size of FF feature vector is reduced to a single measure "m1" by applying Linear Discriminant Analysis (LDA). LDA has already been proposed as a method of information fusion for speech detection in noisy environments in [1]. In that work however, the authors apply LDA to MFCC features.

For a given vector **C** of FF features, the measure m1 is obtained as a scalar product m1=**V**•**C**', where **V** is the eigenvector corresponding to the largest LDA eigenvalue calculated from the training set of parameterized signals. Vectors **C** are of dimension 14 and they are computed from 30 ms long signal frames. The frame shift is 10 ms.

Figure 1 compares visually the discrimination capability of the m1 measure. It shows histograms of m1 values considering the two classes, speech and non-speech, obtained by applying the LDA on FF features (right) and MFCC features (left). The common area shared by the two classes in the FF+LDA

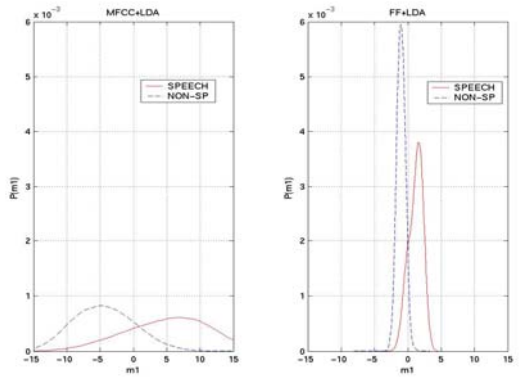

Figure 1 Histograms of m1 values obtained for speech and non-speech classes using MFCC (left) and FF (right) features.

histogram is smaller than that in the MFCC+LDA histogram, indicating the FF+LDA m1 measure is more discriminative than the MFCC+LDA m1 measure. This observation can be confirmed by calculating the classification error of the speech and non-speech m1 measures using an optimal Bayes classifier. Indeed, the classification error for the FF+LDA m1 measure is 10.67%, while for the MFCC+LDA m1 measure it is almost double, 19.81%, for the used training data.

Usually, in speech recognition, delta and delta-delta features are appended to the original static features. These dynamic features, calculated as derivatives of the static ones, carry information about the changes in features along the time and significantly improve the ASR performance. Also in the presented SAD approach, we added to the static FF feature vector of length 14 the corresponding delta and delta-delta features (ΔFF and ΔΔFF) together with the delta energy parameter (ΔE); thus, in total the FF feature vector size increases to 43 (14+14+14+1). After applying LDA to such feature vectors, the classification error for the FF+LDA m1 measure decreases to 7.78%. We use this feature vectors in our SAD system and from now on we refer to the corresponding m1 measure as FF+LDA m1 measure.

3. SPEECH/NON-SPEECH CLASSIFIER AND DECISION BLOCK

The FF+LDA measure m1 is a float number and it has to be post-processed to obtain a binary output indicating speech ("1") and non-speech ("0") portions of signal. The simplest approach is to establish a threshold for making the speech/non-speech decision but this approach gave very poor results. Other approach would be to implement a finite-state automata controlled by the m1 measure levels and to define the time hangover thresholds. However, this approach may require a lot of manual tuning. In the present work, we employed the C4.5 algorithm [3] to train a decision tree classifier with the FF+LDA m1 values (the problem of manual tuning is "left" on the classifier). The output of the tree classifier, the speech/non-speech decision together with its confidence, is passed to Decision Block (see Figure 2). In this block, the binary output is obtained by applying a threshold λ on the confidence.

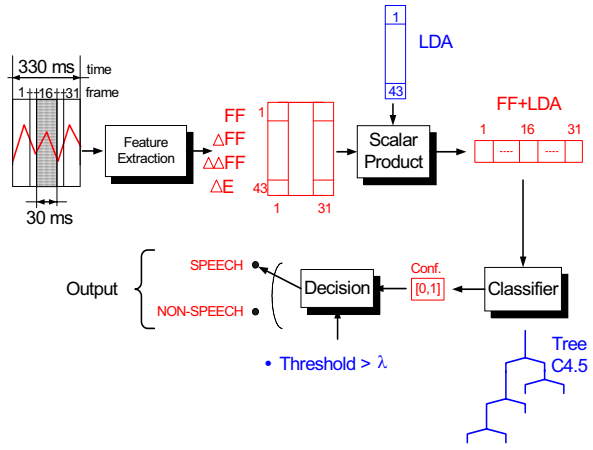

Figure 2 Operational diagram of the proposed SAD.

We tested also neuronal networks as a classifier but the results we obtained were not much different from those of the decision tree classifier; note that much longer training is required for neural networks than for the employed tree classifier.

3.1. Training of the tree classifier

In order to include information from a time span larger than one frame, the classifier is provided also with 15 previous and 15 future m1 values besides the present m1 value (this can be seen as a memory of a finite-state automata). Thus, in total 31 FF+LDA m1 values are available to the classifier, which correspond in our case to a time span of 330 ms.

The classifier was further simplified in such a way that only the most important m1 measures out of the 31 provided were automatically selected. This automatic selection consists in increasing the number of m1 measures allowed to be used by the classifier establishing limits to the growth of the tree: when the decision tree is allowed to use only a single m1 measure, it chooses the present one (the number 16 in Figure 2); in this case, it achieves a 7.78% classification error on the training data. If it is allowed to use two m1 measures, it selects 16 and 6 obtaining an error of 6.1%. Thus, a significant improvement can be achieved by adding only one past m1 measure within about a 100 ms time span (16 less 6 = 10 frames). It is not until the classifier is allowed to choose 4 measures that one selected measure is from the future time with respect to the present measure (the number 27). With seven measures allowed, the classifier chooses – by order of the question within the tree – the measures 16-6-2-27-10-29-5 and it obtains 4.5% error; 4 measures are from past and 2 measures are from future. If the classifier is let to use all 31 measures, the error decreases to 3.6%, however, this does not compensate for the increase in the decision tree size. Thus, we use the previously mentioned seven m1 measures in our SAD system.

In addition, a few experiments have been performed giving an option to the tree classifier to choose the m1 values obtained using the second and higher LDA eigenvector. The tests showed the m1 measures obtained with the first LDA eigenvector from the frames from different time positions are more important than the m1 measures obtained with any other LDA eigenvectors of the present frame.

4. EXPERIMENTS

4.1. Databases

We use two databases in our tests: Spanish SpeechDat [5] and SPEECON [6]. SpeechDat was used for computing the LDA eigenvectors as well as for training the tree classifier. This corpus has been recorded through a fixed telephone line at 8 kHz sampling frequency with the a-law codification. It contains 1011 speakers with 19286 sentences.

All the speech detection experiments were performed on SPEECON data. This database was recorded at 16 kHz sampling frequency and uses 16-bit linear quantification. It contains 600 speakers with both read and spontaneous speech. Interesting for our evaluation is that all the sessions have been recorded simultaneously with four different microphones: a close-talk

microphone, a lavalier microphone, a directional microphone located at 1 m from speaker and an omnidirectional microphone located at 2-3 m from speaker. A subset of 1338 sentences (75 speakers) containing dates, numbers and times of the day were selected for testing. Additionally, in the ASR experiments described bellow, 125/4504 speakers/sentences were used for training the acoustic models. All the used SPEECON data were recorded in the office environment with some occasional computer or air-conditioner noise.

4.2. Metric and reference labels

We use two different kinds of metrics to evaluate our SAD. The first one is a set of general SAD metrics in which the speech/non-speech decision of the evaluated SAD system is compared and scored with the reference speech/non-speech labels. In the other case, the accuracy of an ASR system containing the evaluated SAD system is used as metric for comparison of different SADs.

Although probably the best reference labels for SAD evaluation would be obtained by performing a manual transcription, the generation of such transcription is time consuming and expensive. In this work, we employed a faster and cheaper automatic process using the Viterbi alignment to mark speech and non-speech segments of signal. Note that there is an error related with this labeling. In the case of SPEECON data, the alignment was done on the close-talk microphone and it was adjusted to the other microphones by correlating the respective signals.

4.3. Tested SAD systems

Besides the SAD system described in this paper, three other systems are tested for the comparison purposes. One of them is the commercially used SAD system from the GSM cell-phone standard. The other two are extracted from the ETSI Advanced Front-End standard for noisy speech recognition [4]; one of them is used for frame dropping in this standard (denoted here as AFE_FD) and the other one is used for noise estimation in the de-noising part (Wiener Filter) of the standard (AFE_WF).

4.4. Evaluation by a set of general SAD metrics

We use the following set of general SAD metrics:
1) ***Mismatch Rate*** (***MR***) – gives an average performance of SAD on the data. It is calculated as MR = Time of Incorrect Decisions / Time of All Utterance.
2) ***Speech Detection Error Rate*** (***SDER***) – assesses the SAD performance on the speech portions of signal. It is calculated as: SDER = Time of Incorrect Decisions at Speech Segments / Time of Speech Segments.
3) ***Non-speech Detection Error Rate*** (***NDER***) – assesses how the SAD performs on the non-speech portions of signal. It is calculated as: NDER = Time of Incorrect Decisions at Non-speech Segments / Time of Non-speech Segments.

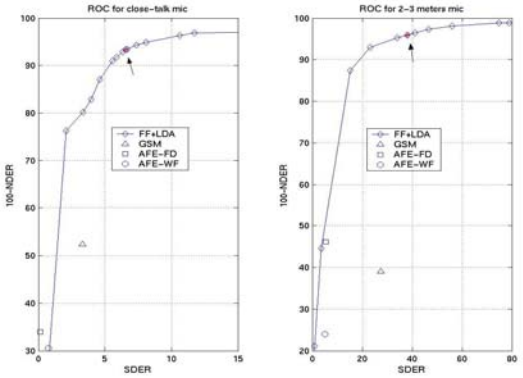

Figure 3 SDER vs. correct non-speech detection. The arrow points the working point with λ=0.5.

Figure 3 shows the relationship between SDER and 100-NDER (= correct non-speech detection) for our FF+LDA SAD system when the threshold λ from Figure 2 changes from 0 to 1.0. Also in Figure 3 there are shown the working positions of the other three SAD systems: GSM, AFE_FD, and AFE_WF. Results for the close-talk and the omni-directional 2-3 m mikes are displayed.

In Figure 3, in GSM SAD the SDER increases from 3.32% observed for the close-talk microphone to 27.49% for the distant microphone and the correct non-speech detection changes from 52.36% to 38.97%. This behavior can be expected from this SAD system as it was designed for close-talk microphone applications. The AFE_FD SAD works better in the distant microphone case. Our SAD system always gives better relation between SDER and correct non-speech detection. For the rest of the experiments we use λ=0.5 in our system.

Table 1 shows detailed SAD results according to the speaker-microphone distance. In general, we observe that all the SADs systems maintain their close-talk MR performance up to the 1 m Directional case. Then, they experiment a significant degradation for the 2-3 m distance except for the AFE_FD SAD. The AFE based systems are very conservative when deciding for non-speech; this can be observed from their low SDER and high NDER obtained for all mikes. On the other hand, the FF+LDA system is more aggressive when deciding for non-speech, which is reflected in relatively balanced SDER and NDER across the first three mikes and the much lower NDER than SDER in the 2-3 m Omni case. It can be foreseen that the AFE systems are well designed for a technique like frame dropping, while the FF+LDA will work better when a good estimation of non-speech is needed (e.g. speech enhancement).

	Close-Talk			Lavalier			1 m Directional			2-3 m Omni		
	SDER	**NDER**	**MR**	**SDER**	**NDER**	**MR**	**SDER**	**NDER**	**MR**	**SDER**	**NDER**	**MR**
GSM	3.32	43.64	18.60	5.83	40.95	17.94	4.61	54.31	21.77	27.49	61.03	39.07
AFE_FD	0.14	66.05	22.86	1.34	49.77	18.04	1.27	57.69	20.75	5.46	53.87	22.17
AFE_WF	0.74	69.46	24.43	3.27	56.52	21.62	2.48	66.55	24.61	4.97	76.01	29.55
FF+LDA	6.57	6.65	6.60	9.78	5.28	8.23	9.47	5.37	8.06	38.12	4.05	26.35

Table 1 Detection errors for the different SADs for different distances from the speaker. In FF+LDA we used λ=0.5.

4.5. Evaluation by ASR accuracy as a metric

For ASR tests we used RAMSES, a speech recognition system developed at our research center. We use 539 demiphoneme units (half of a triphoneme; see [7] for details) modeled by two-state semi-continuous HMMs. The codebook size was 512. The acoustic models were trained with the close-talking SPEECON data.

We used FF features with their first and second derivatives appended as speech parameters for ASR. A simple speech enhancement technique based on Spectral Subtraction (SS) was used to de-noise the signal. The information from SAD is used to update the noise estimate in SS. In addition, other two noise-robust techniques developed at our center were used in combination with SS: the Vector Taylor Series noise compensation (VTS, [8]) and the Quantile based Histogram Equalization technique (QHE, [9]). These noise-robust techniques also benefit from the information provided by the used SAD system.

In the previous tests, the FF+LDA and AFE_FD SAD systems achieved good performance in terms of NDER and SDER, respectively. We test these two SAD systems in our ASR tests. Labeling only the first 4 frames as non-speech was also tested as a simple SAD.

Table 2 shows the word accuracy when employing the FF features with noise-robust techniques and different SAD systems. As a baseline, the results with only FF are reported. We can observe on 2-3 m Omni results that the robustness of the ASR system strongly improves when using the noise-robust techniques supplied by the speech/non-speech information from all SAD systems. The largest improvement is observed for the proposed FF+LDA SAD system. On the other hand, the close-talk mike performance decreases dramatically for the tested systems, except for the FF+LDA SAD system, where a small decrease can be seen.

	Close-Talk	2-3 m Omni
FF		
Baseline	97.17	25.11
FF and SS+VTS+QHE		
4 frames	91.38	53.58
AFE_FD	90.96	55.82
FF+LDA	96.30	60.91

Table 2 Word accuracy (%) for the combination of SS, VTS and QHE techniques using AFE_FD and FF+LDA SAD

5. CONCLUSIONS

In this work, we proposed a robust speech activity detection algorithm based on the Frequency Filtering (FF) features. We illustrated by histograms that the measure obtained by applying Linear Discriminant Analysis (LDA) on FF features is more discriminative in speech – non-speech separation than the same measure based on cepstral features.

We designed a tree classifier which automatically selects the most appropriate FF+LDA measures out of the time sequence of 31 measures covering about 330 ms of signal. During this process we observed the second most useful signal segment for speech/non-speech classification besides the current one is located in about 100 ms before the current signal segment. The importance of this observation is currently under investigation.

Two kinds of metrics were used to evaluate the SAD algorithm: a set of general SAD metrics and the accuracy of an ASR system containing the evaluated SAD system. The tests were done on signals acquired by different microphones, including close-talking and 2-3 m distant microphones. The proposed SAD algorithm shows the best average results among all tested SAD systems for both the set of general SAD metrics and the ASR performance.

5. ACKNOWLEDGMENTS

The authors thank to Juraj Kacur from Slovak University of Technology for his contribution, and also to SIEMENS for allowing the use of the Spanish SPEECON database. This work has been partially sponsored by the European Union under grant IST-2001-28323 (FAME project), and the Spanish Government under grant TIC2002-04447-C02 (ALIADO project).

7. REFERENCES

[1] Martin, A., Charlet, D. and Mauuary, L., "Robust Speech/Non-Speech Detection Using LDA Applied to MFCC", *ICASSP*, 2001.

[2] Nadeu, C., Macho, D., and Hernando, J., "Frequency and Time Filtering of Filter-Bank Energies for Robust HMM Speech Recognition", *Speech Communication*, Vol. 34, pp. 93-114, 2001.

[3] Quinlan, J. R., "C4.5: Programs for Machine Learning", *Morgan Kaufmann*, 1992.

[4] ETSI ES 202 050 Ver. 1.1.1, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced feature extraction algorithm", 2002.

[5] Moreno, A., "SpeechDat Spanish Database for Fixed Telephone Networks. Corpus Design", Technical Report, *SpeechDat Project LE2-4001*, 1997.

[6] Iskra, D. J. et al., "SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation", *Proceedings LREC'2002*, 2002

[7] Mariño, J.B. et al., "The Demiphone: an Efficient Contextual Subword Unit for Continuous Speech Recognition", *Speech Communication, Vol. 32, No. 3, pp. 187-197*, 2000.

[8] Segura, J.C. et al., "VTS Residual Noise Compensation", *ICASSP*, 2002.

[9] Hilger, F. and Ney, H., "Quantile Based Histogram Equalization for Noise Robust Speech Recognition", *EUROSPEECH,* 2001.