Segmentation Experiments for NIST SRE

Technical Report

Jesús Antonio Villalba López

Brno University of Technology

2^{nd} October 2009

1. Introduction

The sites participant in NIST Speaker Recognition Evaluations [1] make use of different Voice Activity Detection (VAD) algorithms for frame selection. The fact, that each site uses different classification approaches and training data makes it difficult to decide which of these algorithms is the best for the speaker verification task. The question arises whether it is better to select all speech frames present in the signal or, on the contrary, keep only the more energetic ones, like voiced speech, that have likely higher discriminative capability. The purpose of this work is to compare the effect of different frame selection approaches on a common Joint Factor Analysis classification framework.

Most of the VAD algorithms used in the NIST SRE can be classified in two groups: energy based (LIA, TNO, Agnitio, I3A) and phone recognition based (BUT, SUNSDV). The energy based VAD are quite similar so we have chosen to use the LIA algorithm to compare with the BUT phone recognizer. Besides, we are going to use Wavesurfer pitch detector to select voiced frames only and the phone recognizer labels for selecting different clusters of frames (voiced, unvoiced, vowels, consonants, etc.). On the other hand, we are going to try to fuse the information of the different segmentations to improve the recognition performance.

This report is organized as follows. Section 2 contains a description of the different frame selection techniques used in this work. Section 3 describes the fusion techniques used to combinate these segmentations. Section 4 describes the experiments we haver carried out and results. Finally, In section 5 some conclusions are presented.

2. Segmentations Description

2.1. JHU08 labels (Baseline)

This is the VAD used by BUT in SRE2008 and JHU08 workshop. Speech/silence segmentation is performed by the Hungarian phoneme recognizer [2,3], where all phoneme classes are linked to 'speech' class. Segments labeled 'speech' or 'silence' are generated, but not merged yet to preserve smaller segments, a post-processing with two rules based on short time energy is applied first:

1. If the average energy in 'speech' segment is 30 dB less than the maximum energy of the utterance, the segment is labeled as silence.
2. If the energy in the other channel is greater than maximum energy minus 3 dB in the processed channel, the segment is also labeled as silence.

After this post-processing, the resulting segments are merged together. Segments shorter than 20 frames are marked as silence. Only speech segments are used. In case of 1-channel files, rule #2 is not applied.

For interview data of SRE08, first, a Wiener filter was applied and new phoneme strings were generated. All phoneme classes were linked to 'speech' class and no further post-processing was done. After that, we took ASR transcripts of the interviewer and removed his/her speech segments from our segmentation files based on time-stamps provided by NIST.

2.2. LIA Three-Gaussian Energy VAD

Segmentation is performed using a three Gaussians model of the speech Energy. GMM model is trained by EM iterations. Posterior probabilities of a frame belonging to each Gaussian are calculated by forward-backward Viterbi realignment. We consider the frame probability of belonging to the speech class as the sum of the posterior probabilities of the two Gaussians of higher energy. The speech probability is thresholded to select speech frames. Finally, a median filter is applied to eliminate speech segments of less than 6 frames and silence segments of less than 4 frames. This kind of VAD and any other similar has been used used in NIST SRE by LIA [4], UWS , I3A and other sites.

2.3. Wavesurfer Pitch Detector

We have used the pitch detector implemented in the open source package Wavesurfer [5] to select voiced frames of speech. Voiced sounds are expected to be more discriminative than unvoiced due to the fact that the resonant frequencies in the vocal tract depend on the physical characteristics of the speaker. The Wavesurfer pitch detector is based on the RAPT algorithm of Talkin [6]. This algorithm is based on the detection of the maximum values of the autocorrelation function in each speech frame. These maxima are fed into a dynamic programming algorithm which decides, for each frame, if it is voiceless or voiced and the more likely pitch frequency. To prevent pitch detection in the cross-talk, segment with energy 30 dB less than the maximum energy of the utterance are eliminated.

2.4. Wavesurfer Voiceless

We select the voiceless frames taking all the JHU08 speech frames that are not selected by the Wavesurfer pitch detector.

2.5. Hungarian Phone Recognizer N1500

Speech/silence segmentation is performed by the Hungarian phoneme recognizer with 1500 neurons, where all phoneme classes are linked to ’speech’ class.

2.6. Hungarian Phone Recognizer N200

Speech/silence segmentation is performed by the Hungarian phoneme recognizer with 200 neurons, where all phoneme classes are linked to ’speech’ class.

2.7. Hungarian Phone Recognizer Clusters

Different phoneme clusters are selected using the string labels of the Hungarian phone recognizer jointly to the SAMPA chart [7]:

- voiced/voiceless.
- vowels/consonants.
- plosives/fricatives/nasals/glides/liquids

Cluster	Units
voiced	A:, b, b:, d, d_, d_:, dz, E, e:, F, g, h1, i, i:, J, J:, j, j:, l, l:, m, m:, N, n, n:, O, o, o:, r, r:, u, u:, v, y, y:, Z, z, z:, :2, _2
voiceless	f, h, k, k:, p, S, S:, s, s:, t, t:, tS, tS_, ts, ts_, t1, t1:, x
vowels	A:, E, e:, i, i:, O, o, o:, u, u:, y, y:, :2, _2
consonants	b, b:, d, d_, d_:, dz, F, f, g, h, h1, J, J:, j, j:, k, k:, l, l:, m, m:, N, n, n:, p, r, r:, S, S:, s, s:, t, t:, tS, tS_, ts, ts_, t1, t1:, v, x, Z, z, z:
plosives	b, b:, d, d_, d_:, dz, g, k, k:, p, t, t:, tS, tS_, ts, ts_, t1, t1:
fricatives	dz, f, h, h1, S, S:, s, s:, tS, tS_, ts, ts_, v, x, Z, z, z:
nasals	F, J, J:, m, m:, N, n, n:
glides	j, j:
liquids	l, l:, r, r:

Table 1. Phoneme clusters.

3. Fusion Description

We have tried to improve the performance of the system carrying out the fusion of systems with different segmentations. Two levels of fusion have been tried: score level fusion and statistics level fusion.

3.1. Score Level Fusion

Score level fusion is based on the combination of the final scores of different systems after ZT-Norm. For that porpoise, logistic regression fusion with the Niko Brummer's FoCal [8] package has been used. The fusion has been done gender dependent. We have tried this fusion only on the dev08 trial list so a cross validation procedure has been used for not training and testing the fusion on the same data:

- We split the trial list in five parts.
- We use four parts for training fusion and we use it for getting the scores of the fifth part. In this way we get scores for the five parts.
- We pool the scores of the five parts and measure the performance as usual.

3.2. Statistics Level Fusion

Statistics Concatenation

This fusion is based on concatenating the zero and first order statistics (N_i,F_i) of each segmentation, jointly with their respective UBM means and variances. After that, we can use this extended statistics to train JFA hyper-parameters (u,v,d) as usual.

$$N=[N_1, N_2, \ldots, N_n]^T$$
$$F=[F_1, F_2, \ldots, F_n]^T$$
$$m=[m_1, m_2, \ldots, m_n]^T$$
$$\Sigma=[\Sigma_1, \Sigma_2, \ldots, \Sigma_n]^T$$

Statistics Weighted Sum

This fusion is based on doing a weighted sum of the zero and first order statistics (N_i,F_i) of each segmentation. This statistics need to be got using a common UBM trained using all the clusters of frames we are going tu fuse.

$$N=\sum_{i=1}^{n} w_i N_i \quad F=\sum_{i=1}^{n} w_i F_i$$

4. Experiments

4.1. Experiments on Dev08 Trial Set

Experiments General Description

We have conducted experiments using de dev08 trial set for the different kind of segmentations and different kind of fusion schemes. These are the parameters common to all the experiments:

- As features we have used 19 MFCC + log Energy augmented with their first and second derivatives getting a 60 dimensional feature vector. The analysis window is 25 ms. length and 10 ms. shift. This features has been short time gaussianized using a 3 s. window.
- Gender Independent UBM GMM have been trained using Mixer 2004 and Mixer 2005 telephone data.
- JFA Hyperparameters (u,v,d) have been trained using a subset of JHU08 training lists corresponding to Mixer 2004 and 2005 signals. A set of 300 eigenvoices and 100 eigenchannels on the speakers with al least 8 recordings (376 females and 294 males). The d matrix describing the remaining speaker variability is trained on top of eigenvoices and eigenchannels. A disjunct set of Mixer 2004 and 2005 speakers with less than 8 recordings (44 females and 13 males) is used for this purpose and MAP estimates of speaker and channel factors are fixed for estimating the diagonal matrix. The Matrices are trained by 10 ML iterations.
- Unless said the contrary, the training of GMM and JFA matrices is conditioned by the selected VAD segmentation.
- Linear scoring for the evaluation of trials has been used in all the experiments.
- Scores are normalized using ZT-Norm. For that, we have used Niko Brummer's lists of speakers Mixer 2004 and 2005 used in JHU08 workshop.

Segmentation Comparative Using All Classes of Speech Frames

In this experiments we compare different segmentations taking all the phone clusters as speech frames. We have used JHU08 labels, LIA energy VAD with different thresholds for the posterior probability of a frame of being speech, Wavesurfer pitch detector and phone recognizer with 1500 and 200 neurons. The EER and DCF results are shown in Table 2. For all the experiments we have used 512 Gaussians.

Segmentation	EER(%)			DCF(%)			% Selected frames related to JHU08
	all	male	female	all	male	female	
JHU08	3,68	3,5	3,87	1,77	1,6	1,86	100
LIA(P(speech)>0.7)	3,62	3,26	3,91	1,88	1,67	2,03	69,42
LIA(P(speech)>0.5)	**3,29**	**3,2**	**3,36**	1,83	1,66	1,9	77,14
LIA(P(speech)>0.3)	3,35	3,21	3,46	**1,62**	**1,52**	**1,68**	81,56
Voiced Wavesurfer	3,65	3,26	3,91	1,85	1,61	1,99	70,14
PHN_HUN_N1500	3,57	3,4	3,66	1,75	1,61	1,83	107
PHN_HUN_N200	3,59	3,45	3,62	1,77	1,6	1,86	116

Table 2. EER and DCF for different VAD.

We can see too there is no much different between using 1500 or 200 neurons in the phone recognizer. LIA VAD get the best results using de 80 % of the frames than JHU08. Despite that we can say there is not big different between then given that LIA(P>0.5) is better in EER but worse than JHU08 in DCF and LIA(P>0.3) is better in DCF than JHU08 but worse in EER than LIA(P>0.5).

Segmentation Comparative Using Voiced/Voiceless Wavesurfer Clusters

In this experiments we compare the results we get using the voice/voiceless clusters given by Wavesurfer and their fusion. In Table 3. we show score fusion and concatenating of stats fusion of voiced unvoiced improve the results of JHU08 even with 1024 Gaussians.

Segmentation	NGauss.	EER(%)			DCF(%)			% Selected frames related to JHU08
		all	male	female	all	male	female	
JHU08	512	3,68	3,5	3,87	1,77	1,6	1,86	100
JHU08	1024	3,4	3,13	3,66	1,77	1,58	1,88	100
Voiced Wavesurfer	512	3,65	3,26	3,91	1,85	1,61	1,99	70,14
Voiceless Wavesurfer	512	7,55	7,91	7,42	3,57	3,34	3,74	29,86
Score Fusion Voiced+Voiceless	2x512	3,15	2,89	3,41	1,72	1,46	1,91	x
Score Fusion Voiced+Voiceless+JHU08	3x512	**3,01**	**2,82**	**3,12**	**1,62**	**1,36**	1,82	x
Stats cat Voiced+Voiceless	2x512	3,29	3,14	3,31	**1,62**	1,4	**1,76**	x

Table 3. EER and DCF for Voiced/Voiceless Wavesurfer Clusters.

Segmentation Comparative Using Voiced/Voiceless Hungarian Phoneme Recogn. Clusters

In this experiments we compare the results we get using the voice/voiceless clusters given by the Hungarian phone recognizer and their fusion. In Table 4. we show the results for the phone recognizer with 1500 neurons. Contrary to the pitch detector clusters we don't get any gain with the fusion unless we add the system using all frames to it.

Segmentation	NGauss.	EER(%)			DCF(%)			% Selected frames related to PHN_HUN_N1500
		all	male	female	all	male	female	
PHN_HUN_N1500	512	3,57	3,4	3,66	1,75	1,61	1,83	100
Voiced	512	3,79	3,76	3,86	1,81	1,74	1,86	77,16
Voiceless	512	11,39	11,47	11,33	5,09	5	5,07	22,84
Score Fusion Voiced+Voiceless	2x512	3,77	3,59	3,91	1,81	1,73	1,86	x
Score Fusion Voiced + Voiceless + PHN_HUN_N1500	3x512	**3,35**	**3,07**	**3,56**	**1,66**	**1,59**	**1,71**	x
Stats cat Voiced+Voiceless	2x512	4,15	4,07	4,1	1,9	1,94	1,87	x

Table 4. EER and DCF for Voiced/Voiceless 1500 Neurons Hung Phon. Recogn. Clusters.

In Table 5. we have the results for 200 neurons recognizer clusters. The number of frames selected as well as the results are quite similar to the 1500 neurons recognizer. In this case we don' t get much gain from the fusion either.

Segmentation	NGauss.	EER(%)			DCF(%)			% Selected frames related to PHN_HUN_N200
		all	male	female	all	male	female	
PHN_HUN_N200	512	3,59	3,45	**3,62**	**1,77**	**1,6**	**1,86**	100
Voiced	512	3,55	**3,14**	3,76	1,8	**1,6**	1,94	76,54
Voiceless	512	11,56	11,73	11,21	5,06	5,09	4,85	23,46
Scr. fusion Voiced+ Voiceless	2x512	**3,51**	3,26	3,76	1,82	1,63	1,93	x

Table 5. EER and DCF for Voiced/Voiceless 200 Neurons Hung Phon. Recog. Clusters.

Segmentation Comparative Using Vowel/Consonant Hung. Phoneme Recogn. Clusters

In this experiments we compare the results we get using the vowel/consonant clusters given by the Hungarian phone recognizer and their fusion. In Table 6. we have the results for the 1500 neurons recognizer.

Using only the vowels that are around half of the frames we can get a result quite near the result using all frames. On the contrary, consonants that have almost the same number of frames get worse results. The consonant clusters of plosives and fricatives have similar error rates and similar number of frames so we can assume they have similar discriminative ability. Nasals have similar error rate to plosives and fricatives but less frames so we can suppose they are more discriminative. Probably this is due to nasals are voiced while plosives and fricatives include voiced and voiceless units. On the other side, glides and liquids have much worse results than the other but have very few frames so we cannot decide if these results are due to less discriminative ability or not having enough frames.

Another thing we have noted is even in the cluster with very few frames, it is better to increase the number of Gaussians from 256 to 512.

The same as for the voiced/voiceless, we don't get any gain of the fusion of clusters unless we add the system with all the frames to the score fusion.

Segmentation	NGauss.	EER(%)			DCF(%)			% Selected frames related to PHN_HUN_N1500
		all	male	female	all	male	female	
PHN_HUN_N1500	512	3,57	3,4	3,66	1,75	1,61	1,83	100
Vowel	512	4,13	3,7	4,4	1,96	1,77	2,04	54,19
Consonants	512	10,95	11,48	10,53	4,3	4,16	4,37	45,81
Plosives	256	12,75	12,35	12,96	5,46	5,08	5,72	15,34
Plosives	512	11,53	10,8	11,92	5,26	4,78	5,61	15,34
Fricatives	256	12,43	12,42	12,77	5,42	5,22	5,42	18,01
Fricatives	512	10,53	9,79	11,12	5,07	4,57	5,15	18,01
Nasals	256	14,3	14,31	14,29	5,46	5,3	5,54	8,34
Nasals	512	12,28	12,61	12,12	4,99	4,97	5	8,34
Liquids	256	22,23	18,7	25,02	8,11	7,47	8,55	3,31
Liquids	512	20,07	17,32	23,59	7,82	7,08	8,32	3,31
Glides	256	23,48	22,65	24,14	8,22	8,09	8,31	2,04
Glides	512	22,34	21,72	22,99	8,1	7,9	8,19	2,04
Score Fusion Vowel + Consonant	2x512	3,87	3,83	4,01	1,87	1,69	1,98	x
Score Fusion Vowel + Plosives + Fricatives + Nasals + Liquids+Glides	512+5x256	3,93	3,58	4,16	1,92	1,76	2,03	x
Score Fusion Vowel + Plosives + Fricatives + Nasals + Liquids+Glides	6x512	3,73	3,39	3,96	1,87	1,71	1,96	x
Score Fusion Vowel + Consonant + PHN_HUN_N1500	3x512	**3,37**	**3,25**	**3,47**	**1,68**	**1,57**	**1,76**	x
Stats cat Fusion Vowel + Consonant	2x512	3,95	4,15	3,66	1,9	1,98	1,83	x

Table 6. EER and DCF for Vow/Cons 1500 Neurons Hung Phon. Recogn. Clusters.

In Table 7 we show the results for the 200 neurons recognizer cluster. Again, the results are quite similar to the ones of the 1500 neurons. The main difference is in the consonants clusters that gets much better results for the 200 neurons clustering. Perhaps, for this recognizer the limits between clusters are not so well defined and some vowels frames are selected as consonants.

Segmentation	NGauss.	EER(%)			DCF(%)			% Selected frames related to PHN_HUN_N200
		all	male	female	all	male	female	
PHN_HUN_N200	512	**3,59**	**3,45**	**3,62**	**1,77**	**1,6**	**1,86**	100
Vowel	512	4,34	4,02	4,65	2,06	1,88	2,13	51,98
Consonants	512	5,56	5,14	5,88	2,79	2,44	3,01	48,02
Plosives	512	12,75	12,37	13	5,72	5,38	5,93	13,73
Fricatives	512	10,63	10,04	10,73	4,73	4,63	4,72	20,15
Nasals	512	12,56	12,3	12,81	5,04	4,92	5,1	8,19
Liquids	512	16,27	14,18	17,96	7,03	6,07	7,49	4,48
Glides	512	20,08	19,3	20,64	7,46	7,46	7,54	2,41

Table 7. EER and DCF for Vow/Cons 200 Neurons Hung Phon. Recogn. Clusters.

Weighted Sum of Statistics Fusion

In this section we show the results for the fusion based on doing a weighted sum of the statistics calculated using different segmentations. Unlike the other experiments, for this fusion all the statistics need to be calculated using the same GMM. For that porpoise, we have used the UBM trained using the phone recognizer VAD because it includes all kinds of frames.

We have selected the weights in such manner as the frames supposed to be more discriminative have weight and the others, less than one. In Table 8, we show the results. We have got some improvement fusion LIA and phone recognizer labels but not in the others. We tried to estimate the optimum value for the weights using a discriminative approach but some approximations were needed for making the equations affordable and we didn't manage to make the algorithm converge.

Segmentation	NGauss.	EER(%)			DCF(%)		
		all	male	female	all	male	female
PHN_HUN_N1500	512	3,57	3,4	3,66	1,75	1,61	**1,83**
LIA05	512	**3,29**	3,2	**3,36**	1,83	1,66	1,9
PHN_HUN_N1500(0.25) + LIA05(0.25) + LIA03(0.25) + Voiced(0.25)	512	3,51	3,07	3,76	1,71	1,51	1,86
PHN_HUN_N1500(0.3) + LIA05(0.7)	512	3,4	**3,01**	3,67	**1,68**	**1,47**	**1,83**
Voiced(1) + Unvoiced(0.3)	512	3,84	3,64	4,1	1,84	1,67	1,94
Vow(1) + Cons(0.3)	512	4,09	3,69	4,3	1,65	1,76	2,07

Table 8. EER and DCF for weighted sum of statistics fusion.

4.2 Experiments on SRE08 short2-short3 condition

We have repeated some experiments on the SRE08 trial sets to check if the results with the dev08 trials sets holds and specially to see how different VAD behave in the microphone conditions. We show the results in Tables from 9, 10 and 11.

The configuration of the system is almost the same as in the dev08 tests. The only difference is another set of 100 eigenchannels is trainned on 2005 microphone data (52 females and 45 males). Both matrices (phone and microphone) are stacked together having 200 eigenchannels in total.

In the SRE08 interview signals, speech of the target speaker and the interviewer is mixed in the same channel. We have run the experiments leaving the interviewer speech and using the ASR transcripts provided by NIST to eliminate it. Despite the interviewer represents only around a 3% of the speech frames we get an interesting improvement.

The labels of JHU08 are the best for the microphone conditions. The main difference between JHU08 and PHN_HUN_N1500 is the preprocessing of the speech with a Wiener filter. It seems the filter helps to get better phone labels. Leaving JHU08 labels out, the PHN_HUN_N1500 gets better performance than LIA en most of cases. Wavesurfer fails completely detecting pitch in some microphone files which implies very poor performance due to some speaker models are not trained at all.

Besides, we have tried use two different VAD to estimate speaker and channel. We have used LIA VAD that, in theory, should produce more discriminative frames for estimating speaker and common factors and Hungarian phone recognizer to estimate channel factors. We have done that only in training or in training and test. In both cases, the performance is worse.

Segmentation	EER(%)			DCF(%)			% Selected frames related to JHU08
	det1	det2	det3	det1	det2	det3	
JHU08	**8,11**	**1,04**	**8,14**	**4,16**	**0,53**	**4,23**	100
LIA(P(speech)>0.5)	11,49	2,38	11,59	5,73	1,22	5,88	75
Voiced Wavesurfer	16,93	17,32	16,9	6,99	3,26	7,16	81
PHN_HUN_N1500	10,57	1,9	10,44	4,59	0,74	4,64	85
LIA(P(speech)>0.5) without interviewer	9,91	2,25	9,86	4,99	1,22	5,01	72
Voiced Wavesurfer without interviewer	15,5	16,97	15,42	6,21	2,99	6,36	78
PHN_HUN_N1500 without interviewer	9,45	2,03	9,31	4,31	0,72	4,32	82
spk. factors → LIA ch. factors → PHN_HUN_N1500 Train	10,51	2,42	10,76	5,05	0,96	5,19	x
spk. factors → LIA ch. factors → PHN_HUN_N1500 Train-Test	11,36	3,25	11,38	5,68	1,74	5,74	x

Table 9. EER and DCF for SRE08 Microphone Conditions.

Segmentation	EER(%)		DCF(%)		% Selected frames related to JHU08
	det4	det5	det4	det5	
JHU08	**7,77**	8,62	**2,92**	3,22	100
LIA(P(speech)>0.5)	10,43	9,5	4,01	3,35	82
Voiced Wavesurfer	12,22	14,54	4,18	4,19	82
PHN_HUN_N1500	9,59	**7,47**	3,85	**2,89**	102
LIA(P(speech)>0.5) without interviewer	9,79	9,5	3,75	3,35	79
Voiced Wavesurfer without interviewer	11,04	14,54	3,95	4,19	79
PHN_HUN_N1500 without interviewer	8,96	**7,47**	3,59	**2,89**	99

Table 10. EER and DCF for SRE08 Cross Channel Conditions.

In the telephone only conditions all the VAD performs quite similar not being clear which of then is better.

Segmentation	EER(%)			DCF(%)			% Selected frames related to JHU08
	det6	det7	det8	det6	det7	det8	
JHU08	7,17	3,01	3	4,13	1,5	1,46	100
LIA(P(speech)>0.5)	6,98	**3,18**	**3,29**	**3,83**	1,47	1,53	86
Voiced Wavesurfer	7,1	3,5	3,62	4	**1,46**	**1,43**	80
PHN_HUN_N1500	**6,94**	3,33	3,47	4,11	1,5	1,49	116

Table 11. EER and DCF for SRE08 Telephone Conditions.

4.3 Experiments on SRE08 10sec-10sec condition

We have done some experiments with the SRE08 10sec-10sec condition. The configuration of the system is the same as for the long utterances but without the microphone eigenchannel matrix due to the fact that there is no microphone test, and without the d matrix.

We found out our results being very poor compared to other sites participant in the evaluation. After some research we decided to try the eigenchannel integration scoring [9]. This type of scoring assumes that speaker factors have a normal distribution with zero mean and unit variance. Maximum likelihood estimation takes too many iterations to produce this distribution so we re-estimated the hyperparameters by the minimum divergence approach (MD) [10]. With this we found a great improvement not only in the eigenchannel integration scoring but as well in the linear scoring. This confirms that for short utterances a correct estimation of the eigenvectors magnitude is essential for good performance.

As far as the VAD is concerned, the results confirm what we have seen in previous tests. There is no a big difference between segmentations but the phone recognizer works somewhat better. It is our feeling that for long utterances it does not matter to discard some frames but in short utterances it is necessary to keep as much speech frames as possible.

Segmentation	EER(%)			DCF(%)			% Selected frames related to PHN_HUN_N1500
	det6	det7	det8	det6	det7	det8	
PHN_HUN_N1500	27,72	25,07	26	9,33	8,67	8,89	100
LIA(P(speech)>0.5)	28,6	25,94	26,29	9,65	9,09	9,21	73
Voiced Wavesurfer	28,4	25,21	25,14	9,55	8,8	8,89	75
PHN_HUN_N1500 with MD Eigenchannel integration	20,24	17	17,36	**8,1**	7,01	**6,98**	100
PHN_HUN_N1500 with MD	**19,84**	**16,71**	17,71	8,15	**7**	7,03	100
LIA(P(speech)>0.5) with MD	20,92	17,46	18,86	8,4	7,61	7,63	73
Voiced Wavesurfer with MD	20,65	16,86	**16,84**	8,34	7,41	7,61	75

Table 12. EER and DCF of SRE08 10sec-10sec condition.

4.4 Experiments on SRE08 short2-10sec condition

We have done some experiments with the SRE08 short2-10sec condition, too. The configuration of the system is the same as for the long utterances but without the microphone eigenchannel matrix due to the fact that there is no microphone tests. Again there is not much difference between VAD and MD reestimation of the JFA matrices gives an important improvement. Results are shown in Table 13.

Segmentation	EER(%)			DCF(%)			% Selected frames related to PHN_HUN_N1500
	det6	det7	det8	det6	det7	det8	
PHN_HUN_N1500	15,7	11,81	13,06	6,26	4,76	5,33	100
LIA(P(speech)>0.5)	15,97	12,8	13,89	6,63	5,04	5,41	73
Voiced Wavesurfer	15,79	12,96	13,06	6,47	4,96	5,5	76
PHN_HUN_N1500 with MD	**13,32**	**9,42**	10,83	**5,79**	**4,24**	4,86	100
LIA(P(speech)>0.5) with MD	13,32	9,87	**10,23**	6,02	4,32	**4,59**	73
Voiced Wavesurfer with MD	13,44	10,09	10,61	6	**4,24**	4,62	76

Table 13. EER and DCF of SRE08 short2-10sec condition.

4.5 Experiments on SRE08 short2-short3 condition with MD

After getting an interesting improvement using MD re-estimation of the JFA matrices in the 10 sec conditions, we decided to try it in the short2-short3 condition too. The system configuration is the same as in 4.2 without d matrix. We have got some improvement too in most of the conditions but not as much as in 10 sec conditions. As conclusion, we can say that MD doesn't hurt performance in long utterances and it is necessary to assure good error rates across conditions.

Training u,v	EER(%)								DCF(%)							
	det1	det2	det3	det4	det5	det6	det7	det8	det1	det2	det3	det4	det5	det6	det7	det8
ML	9,17	**1,39**	9,34	9,03	7,4	7,02	3,58	**3,38**	4,09	0,45	4,19	3,88	3,28	3,89	1,8	1,88
ML+MD	**8,36**	1,56	**8,51**	**7,69**	**7,07**	**6,8**	**3,42**	3,41	**3,76**	**0,45**	**3,82**	**3,69**	**3,12**	**3,83**	**1,72**	**1,78**

Table 14. EER and DCF of SRE08 short2-short3 condition with ML and ML+MD

4.6 Final Experiments

To end our work we are going to set up the best JFA system as possible. For that, we have selected the next configuration:

- The frame selection is done with the Hungarian phone recognizer of 1500 neurons that looks the more robust across conditions.
- Gender Independent UBM GMM have been trained using Mixer 2004 and Mixer 2005 and Switchboard telephone data.
- JFA Hyperparameters (u,v,d) have been trained using two different sets of lists:
  - JHU08 training lists: A set of 300 eigenvoices is trained on Switchboard, Mixer04 and Mixer05 telephone data on the speakers with 8 recordings. 100 eigenchannels are trained on Mixer04 and Mixer05 telephone data on the speakers with at least 8 recordings. The d matrix describing the remaining speaker variability is trained on top of eigenvoices and eigenchannels. A disjunct set of Mixer 2004 and 2005 speakers with less than 8 recordings (44 females and 13 males) is used for this purpose and MAP estimates of speaker and channel factors are fixed for estimating the diagonal matrix.
  - Similar to Patrick Kenny's training list: A set of 300 eigenvoices is trained on Switchboard and Mixer05 telephone data. 100 eigenchannels are trained on Switchboard, Mixer04 and Mixer05 telephone data. The d matrix is training using Mixer04.
  - 100 eigenchannels are trained using Mixer05 microphone data.
  - The matrices are trained by 10 ML and 10 MD iterations.
- The rest of parameters is the same as in previous experiments.

The results are shown in Tables15 to 17. We compare the results with the ones we got with the previous set up: 512 Gaussians, gender independent UBM and Mixer only data for training UBM and JFA Matrices.

Set up	EER(%)								DCF(%)							
	det1	det2	det3	det4	det5	det6	det7	det8	det1	det2	det3	det4	det5	det6	det7	det8
JHU08 Mixer 512G GI	8,36	**1,56**	8,51	7,69	7,07	6,8	3,42	3,41	**3,76**	**0,45**	**3,82**	3,69	3,12	3,83	1,72	1,78
JHU08 2048G GD	**8,25**	1,9	**8,2**	**6,79**	**5,3**	**5,34**	**2,69**	**2,63**	3,87	0,86	3,88	**2,89**	**2,43**	**3,06**	**1,31**	1,3
Kenny's 2048G GD	8,73	**1,56**	8,82	7,6	6,12	5,9	2,85	**2,63**	3,96	0,65	4,04	3,15	2,44	3,19	1,31	**1,27**

Table 15. EER and DCF of SRE08 short2-short3 condition with different training list for u,v,d.

Set up	EER(%)			DCF(%)		
	det6	det7	det8	det6	det7	det8
JHU08 Mixer 512G GI	19,84	16,71	17,71	8,15	7	7,03
JHU08 2048G GD	**15,56**	**12,39**	**12,63**	**6,89**	**5,62**	**5,79**
Kenny's 2048G GD	16,72	13,54	13,71	7,57	6,39	6,33

Table 16. EER and DCF of SRE08 10sec-10sec condition with different training list for u,v,d.

Set up	EER(%)			DCF(%)		
	det6	det7	det8	det6	det7	det8
JHU08 Mixer 512G GI	13,32	9,42	10,83	5,79	4,24	4,86
JHU08 2048G GD	**10,4**	**6,88**	**7,17**	**4,98**	**3,1**	**3,27**
Kenny's 2048G GD	12,06	8,85	9,22	5,48	3,89	4,28

Table 17. EER and DCF of SRE08 short2-10sec condition with different training list for u,v,d.

We have got an important improvement in all the phone conditions with the new configuration. On the contrary, for microphone conditions there is no gain. The reason can be the system is too telephone conditioned and more speech is needed for training the microphone eigenchannel matrix. Besides, we have got better results with the JHU08 training list than with Kenny's ones. The explanation can be channels in Mixer are more similar to SRE08 channels than channels in Switchboard.

5. Conclusions

In this work he have tested the performance of several VAD algorithms on a JFA speaker recognition framework on dev08 and SRE08 trial sets. We have found that JFA is quite robust to the frame selection approach getting similar results using segmentations that select all kind of speech and segmentations that only select the more energetic frames like LIA. For the telephone conditions we found some improvement using LIA VAD in dev08 but not in SRE08. For the microphone and 10 seconds conditions of SRE08 the phone recognizer VAD seems the best option.

We have done some experiments using different clusters of frames to decide which of them are more discriminative. As expected, we have found that voiced are more discriminative than unvoiced and vowels than consonants. The clusters of consonants given by plosives, fricatives and nasals have similar performance but it is difficult to quantify the discriminative ability of each of then due to each cluster have different number of frames. Nasals look more discriminative because it gets similar results with half the number of frames. Glides and Liquids have very poor performance but they have too few frames for getting any conclusion. Some fusion experiments have been carried out with this clusters but we have got little gain on this. Only when adding the system using all kinds of frames to the fusion with the systems using clusters of frames, we get a clear improvement.

In our experiments with the SRE08 10sec conditions we have found Minimum Divergence re-estimation of the JFA hyperparameters to be essential for good performance. The reason is that a bad estimation of the space variability doesn't allow the speaker model or the channel to adapt enough to the signal. For long conditions can be helpful to but not so much.

Finally, we have set up a system whose performance matches the one of the best systems in SRE08 NIST evaluation. For that, we have selected a gender dependent configuration using 2048 Gaussians and trained with Mixer and Switchboard databases. We have got a great improvement over previous experiments in telephone conditions but not in microphone ones. More improvement can be expected if more microphone data is added to the training of the channel compensation.

Bibliography

[1] The NIST year 2008 speaker recognition evaluation plan. [Online]. Available: http://www.nist.gov/speech/tests/spk/2008/index.htm

[2] Schwarz P., Matejka P. and Cernocky J.: Hierarchical Structures of Neural Networks for Phoneme Recognition, In Proceedings of ICASSP 2006,May 2006, Toulouse, France.

[3] Matejka P., Burget L., Schwarz P. and Cernocky J., Brno University of Technology System for NIST 2005 Language Recognition Evaluation. Odyssey: The Speaker and Language RecognitionWorkshop, San Juan, Puerto Rico, June 2006.

[4] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve and J. Mason, " ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition," in Speaker Odyssey, South Africa, January 2008.

[5] http://www.speech.kth.se/wavesurfer

[6] Talkin, D., 1995. A robust algorithm for pitch tracking (RAPT). In: Speech Coding and Synthesis. Elsevier Science, Amsterdam, NL, pp. 495–518.

[7] http://en.wikipedia.org/wiki/SAMPA_chart

[8] http://www.dsp.sun.ac.za/~nbrummer/focal

[9] Glembek, O., Burget, L., Dehak, N., Brummer, N., and Kenny, P., "Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis", to appear in Proc ICASSP 2009, Taipei, Taiwan, April 2009.

[10] Kenny, P Joint factor analysis of speaker and session variability : Theory and algorithms - Technical report CRIM-06/08-13 Montreal, CRIM, 2005.