Cepstral Trajectories in Linguistic Units for TextIndependent Speaker Recognition

Javier Franco-Pedroso, Fernando Espinoza-Cuadros and Joaquin Gonzalez-Rodriguez

ATVS – Biometric Recognition Group
Universidad Autonoma de Madrid, Spain
javier.franco@uam.es

Abstract. In this paper, the contributions of different linguistic units to the speaker recognition task are explored by means of temporal trajectories of their MFCC features. Inspired by successful work in forensic speaker identification, we extend the approach based on temporal contours of formant frequencies in linguistic units to design a fully automatic system that puts together both forensic and automatic speaker recognition worlds. The combination of MFCC features and unit-dependent trajectories provides a powerful tool to extract individualizing information. At a fine-grained level, we provide a calibrated likelihood ratio per linguistic unit under analysis (extremely useful in applications such as forensics), and at a coarse-grained level, we combine the individual contributions of the different units to obtain a highly discriminative single system. This approach has been tested with NIST SRE 2006 datasets and protocols, consisting of 9,720 trials from 219 male speakers for the 1side-1side English-only task, and development data being extracted from 367 male speakers from 1,808 conversations from NIST SRE 2004 and 2005 datasets.

Keywords: automatic speaker recognition, forensic speaker identification, temporal contours, linguistic units, cepstral trajectories.

1 Introduction[1]

Automatic speaker recognition has focused in the last decade on two concurrent problems: the compensation of session variability effects, mainly through highdimensional supervectors and latent variable analysis [2] [7] [8], and the production of an application-independent calibrated likelihood ratio per speaker recognition trial [1], able to elicit useful speaker identity information to the final user with any given application prior. The results are highly efficient text-independent systems in controlled conditions, as NIST SRE evaluations, where lots of data from hundreds of speakers in similar conditions are available. Thus, all the speech available

[1] Supported by MEC grant PR-2010-123, MICINN project TEC09-14179, ForBayes project CCG10UAM/TIC-5792 and Catedra UAM-Telefonica. Thanks to ICSI (Berkeley, CA) for hosting the preliminary part of this work. Thanks to SRI for providing Decipher labels for SRE datasets.

adfa, p. 1, 2011.

in every trial is used to produce detection performances difficult to imagine a decade ago.

However, in the presence of strong mismatch (as e.g. in forensic conditions, where acoustic and noise mismatch, apart from highly different emotional contexts, speaker roles or health/intoxication states can be present between the control and questioned speech), those acoustic/spectral systems could be unusable as all our knowledge about the two speech samples is deposited into a single likelihood ratio, obtained from all the available speech in the utterance, that could be strongly miscalibrated (being then highly misleading) as the system has been developed under severe database mismatch between training and testing data. Moreover, it is difficult (or even impossible) to collect enough data to develop a system robust to every combination of mismatch factors present in actual case data, an important problem in real applications.

A usual procedure in forensic laboratories is that a speech expert, typically a linguist/phonetician, can isolate or mark segments of compatible/comparable speech between both samples, segments being from seconds long to just some short phonetic events in given articulatory contexts. The number and types of comparable units for analysis is always a case-dependent subject, and therefore flexible strategies for analysis and combination are needed.

The proposed approach gives an answer to this application framework, providing informative calibrated likelihood ratios for every linguistic unit under analysis. Moreover, the combination of the different units yields good discrimination capabilities allowing to obtain speaker detection performance levels similar to equivalent acoustic/spectral systems when enough usable units are available.

The remainder of the paper is organized as follows. In Sections 2 and 3 we present, respectively, our proposed front-end for feature extraction over linguistic units and the system in use. Section 4 describes the databases and the experimental protocol used for testing the system. Section 5 shows results for the different linguistic units individually and for several combination methods, to finally conclude in Section 6 summarizing the main contributions and future extensions of this work.

2 Cepstral trajectories extraction from linguistic units

Many attempts have been made to incorporate the temporal dynamics of speech into features, from the simplest use of the velocity (delta) and acceleration (delta-delta) derivative coefficients to modulation spectrograms, frequency modulation features or even TDCT (temporal DCT) features (see [9] for a review). However, to the best of our knowledge none of the previous approaches, with the exception of SNERFs [4] and [12] for prosodic information, take advantage of the linguistic knowledge provided by an automatic speech recognizer (ASR) to extract non-uniform-length sequences of spectral vectors to be converted into constant-size feature vectors characterizing the spectro-temporal information in a given linguistic unit. In our proposed front-end, we obtain a constant-size feature vector from non-uniform-length MFCC features sequence within a phone unit.

2.1 ASR region conditioning

In this work, both phone and diphone units have been used for defining time intervals in order to extract the temporal contours over the MFCC features. For this purpose, the phonetic transcription labels produced by SRI’s Decipher conversational telephone speech recognition system [6] were used first. For this system, trained on English data, the Word Error Rate (WER) of native and nonnative speakers on transcribed parts of the Mixer corpus, similar to NIST SRE databases used for this work, was 23.0% and 36.1% respectively. These labels define both phonetic content and time interval of speech regions containing the phone units to be segmented. For this work, 41 phone units from an English lexicon were used, represented by the Arpabet phonetic transcription code [13]. Diphone units are defined by the combination of any two consecutive phone units, although only a subset of 98 diphones of the possible combinations was used (those presenting higher frequency of occurrence).

2.2 Cepstral trajectories parameterization

By means of SRI’s Decipher phone labels, trajectories (i.e., the temporal evolution of each MFCC vector dimension) of 19 static MFCC are extracted from phone and diphone units, yielding a MFCC matrix of 19 coefficients x #frames/unit for each linguistic unit. This variable-length segment is duration equalized to a number of frames equivalent to 250 ms. Finally, those trajectories are coded by means of a fifth order discrete cosine transform (DCT), yielding our final 19 x 5 fixed-dimension feature vector for each linguistic unit.

3 System description

3.1 Unit-dependent acoustic systems

Proposed systems are based on the well known GMM-UBM framework [11], using duration-equalized DCT-coded MFCC trajectories per linguistic unit as feature vectors. The GMM-UBM systems have been the state-of-the-art in the text-independent speaker recognition field for many years until the emergence of JFA [7] and total variability [2] techniques, which have outperformed the former ones through accurately modeling the existing variability in the supervector feature space. For this work, GMM-UBM systems have been chosen for two main reasons: i) as we are using a new type of features, we need first to find the optimal configuration for this GMMUBM new-framework, which is the basis of supervector-based systems; and ii) because we aim to model speakers in a unit-dependent way, a much smaller amount of data is available for training purposes, so probably not enough data would be available to capture the existing variability in each unit domain (also having into account that we only have ASR labels from the SRE04, SRE05 and SRE06 datasets).

Three different unit-dependent GMM-UBM configurations were tested previously to perform experiments reported in this paper:

1. UBM and speaker models trained on unit-independent data; evaluation trials performed on unit-dependent test data (as we did in our first approach [5]).
2. UBM trained on unit-independent data; speaker models adapted from unitdependent training data; evaluation trials performed on unit-dependent test data.
3. UBM and speaker models trained on unit-dependent data; evaluation trials performed on unit-dependent test data (fully unit-dependent).

For each configuration, different numbers of mixtures were tested, ranging from 2 up to 1024 mixtures increasing in powers of 2. It was found out that best results were obtained for the fully unit-dependent configuration, using 8 mixtures in the case of phone units and 4 mixtures in the case of diphone units. These configurations are those used to obtain the individual linguistic unit results reported in this paper.

3.2 Fusion schemes and linguistic units combinations

Both individual unit performance and different unit combinations have been analyzed in this paper. On the one hand, individual linguistic-unit systems allow us to report useful speaker verification LR's for very short speech samples where usual state-ofthe-art systems are not directly applicable (as it is the case of forensic applications). On the other hand, when more data is available, individual units can be combined to achieve better discriminative capabilities.

In addiction to obtaining test results for each linguistic unit, these individual systems were combined in both intra- and inter-unit manners, i.e. fusing phone/diphone units between them and fusing phone and diphone units together. Two different fusion techniques were used: sum fusion and logistic regression fusion. The former one was performed after linear logistic regression calibration, while the latter one was performed in a single calibration/fusion step.

Another issue is what should be the selected units to be fused. Two strategies have been used in this work. The first of them is to select the n-best performing units by setting a threshold for the EER of the units to be fused, leaving out those performing worse. However, this procedure do not guaranty that the best fused system will be achieved because some units with lower performance by itself could contribute to the fused system if its LR's are sufficiently low correlated with those produced by the other units to be fused. On the other hand, testing all of the possible combinations would be a very complex task, so we used a unit selection algorithm (similar to that used in [3]) based on the following steps:

1. Take the best performing unit in terms of EER as the initial units set.
2. Take the next best performing unit and fuse with the previous set. If the fusion improves the performance of the previous set, this unit is added to the units set, otherwise rejected.
3. The previous step is repeated for all the units in increasing EER order.

This procedure allows us to find complementarities between units that otherwise would not have been revealed, but avoiding the complex task of testing each possible combination.

4 Datasets and experimental setup

NIST SRE datasets and protocols have been used to develop and test our proposed system, in particular those of years 2004, 2005 and 2006. As region conditioning for linguistic units definition and extraction rely on SRI's Decipher ASR system (trained on English data), English-only subsets of the NIST SRE datasets have been used. SRE 2004 and 2005 datasets were used as the background dataset for UBM training, consisting of 367 male speakers from 1,808 conversations (only male speakers were used for this work). English-only male 1side-1side task from SRE 2006 was used for testing purposes. This dataset and evaluation protocol comprises both native and nonnative speakers across 9,720 same-sex different-telephone-number trials from 298 male speakers. SRE 2005 evaluation set was also used to obtain scores in order to train the calibration rule (linear logistic regression).

Performance evaluation metrics used are the Equal Error Rate (EER) and the Detection Cost Function (DCF) as defined in the NIST SRE 2006 evaluation plan [10]. Cllr and minCllr [1] (and its difference, calibration loss) are also used to evaluate the goodness of the different detectors after the calibration process.

5 Results

5.1 Reference system performance

As we are using the GMM-UBM framework to model unit-dependent systems, our baseline reference system is also a GMM-UBM system based on MFCC features. A classical configuration with 1024 mixtures and diagonal covariance matrices was used, and MFCC features include 19 static coefficients plus first order derivatives, cepstral mean normalization, RASTA filtering and feature warping. The performance of this system in the English-only male 1side-1side task from SRE 2006 is EER=10.26% and minDCF=0.0457. This system does not include any type of score normalization.

5.2 Phone units: individual and combined systems performances

Table 1 shows individual performance of phone units for the NIST SRE 2006 English-only male 1side-1side task. It can be seen that, although most of the phones have high EER and minDCF values, almost all of them are well calibrated (low difference between Cllr and minCllr). This allows us to obtain informative calibrated likelihood ratios from very short speech samples (as low as some phone units), as we can see in the tippet plot in Figure 1 for the best performing phone unit (‘N’). Moreover, there are lots of units that can be combined, and despite their lower individual performance (around 60% worse than the reference system for the best performing phone), combined system can outperform reference system by means of sum or logistic regression fusion, as it can be seen in Figure 2. This is due to the highly complementarity of acoustic systems coming from different linguistic content.

Phone unit	EER (%)	minDCF	C_{llr}	$minC_{llr}$
AA	32.20	0.0983	0.8633	0.8452
AE	18.98	0.0813	0.6087	0.5832
AH	29.39	0.0969	0.8235	0.7967
AO	34.36	0.0992	0.9065	0.8838
AW	36.99	0.0991	0.9241	0.9111
AX	27.08	0.0947	0.7882	0.7512
AY	21.68	0.0869	0.6822	0.6428
B	34.50	0.0986	0.8922	0.8778
CH	42.59	0.1000	0.9686	0.9538
D	32.07	0.0965	0.8661	0.8500

DH	28.43	0.0934	0.8403	0.7857
DX	40.44	0.0998	0.9670	0.9484
EH	31.69	0.0975	0.8574	0.8283
ER	35.18	0.0987	0.9107	0.8901
EY	26.40	0.0925	0.7713	0.7515
F	39.63	0.0993	0.9561	0.9397
G	35.71	0.1000	0.9291	0.9040
HH	39.80	0.0992	0.9527	0.9414
IH	26.95	0.0948	0.7964	0.7495
IY	23.32	0.0923	0.7453	0.7002
JH	39.69	0.0997	0.9487	0.9339
K	27.76	0.0961	0.8219	0.7832
L	26.51	0.0935	0.7789	0.7451
M	22.28	0.0857	0.6824	0.6583
N	15.92	0.0713	0.5520	0.5082
NG	29.37	0.0934	0.9977	0.7958
OW	24.65	0.0987	0.7917	0.7396
P	39.50	0.0988	0.9466	0.9335
PUH	24.18	0.0908	0.7359	0.7149
PUM	34.15	0.0953	0.8644	0.8419
R	24.65	0.0887	0.7295	0.7116
S	30.04	0.0973	0.8451	0.8059
SH	39.36	0.0996	1.0546	0.9294
T	27.89	0.0921	0.8256	0.7647
TH	38.37	0.1000	1.1207	0.9298
UH	41.53	0.1000	0.9717	0.9593
UW	24.79	0.0898	0.7391	0.7198
V	35.86	0.0990	0.9093	0.8932
W	35.82	0.0993	0.9167	0.8966
Y	24.00	0.0906	0.7313	0.7062
Z	32.07	0.0968	0.8487	0.8312

Table 1. EER (%), minDCF, Cllr and minCllr for phone units in the NIST SRE 2006 Englishonly male 1side-1side task.

It should be noted that results equivalent to that of the reference system can be achieved by combining only 4 phone units (‘AE’, ‘AY’, ‘M’, ‘N’). Also, it can be seen that the unit selection algorithm used can achieve better fusion results than simply setting a threshold for the EER of the units to be fused, both for sum and logistic regression fusions. Furthermore, it is worth noting that some of the phone units selected to be fused have very low performance (‘CH’ in the sum fusion, ‘AO’ in both sum and logistic regression fusions).

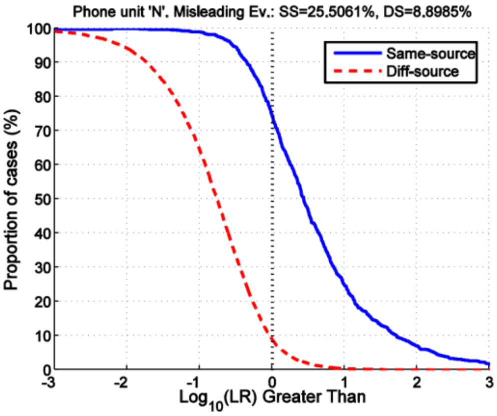

Fig. 1. Tippet plot for the best performing phone unit ('N') in the NIST SRE 2006 English-only male 1side-1side task.

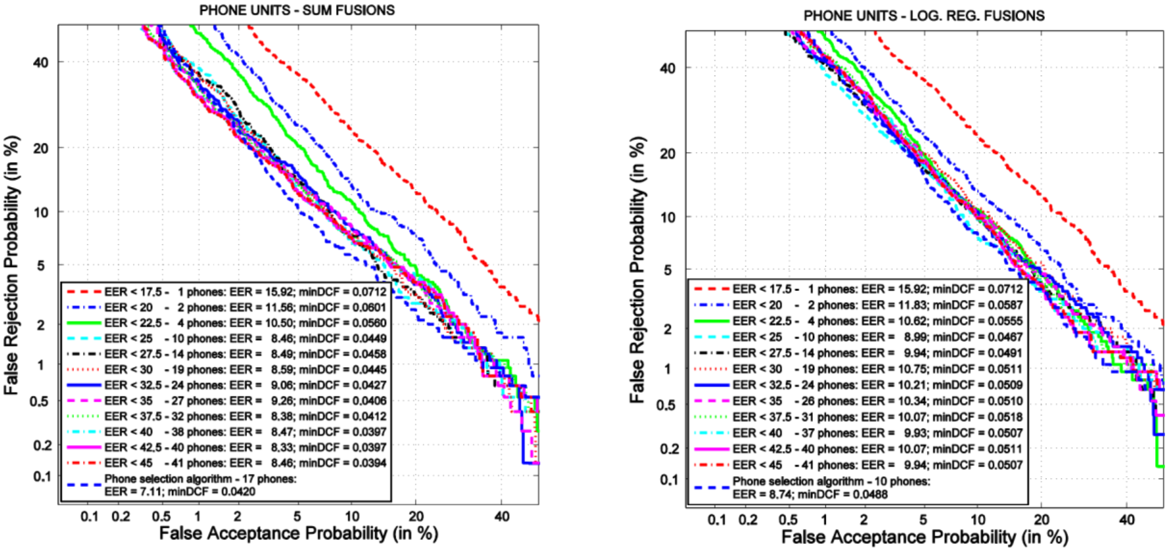

Fig. 2. DET curves for sum and log. reg. fused systems in the NIST SRE 2006 English-only male 1side-1side task for different phone selection schemes.

5.3 Diphone units: individual and combined systems performances

Table 2 shows individual performance for the ten best performing diphone units for the NIST SRE 2006 English-only male 1side-1side task. As it can be seen, diphone units have much lower performance than phone units. This may be due to the fact that, while diphones cover a longer time span that can present more complex trajectories, we are still using a 5 order DCT to code these trajectories. However, as it can be seen in Figures 3, diphone fusions can achieve as good performance as the phones unit fusions, although more units are needed to be fused.

Diphone unit	EER (%)	minDCF	C_{llr}	minC_{llr}

AEN	30.72	0.0993	0.8479	0.823
AET	31.89	0.0969	0.872	0.8526
AXN	23.84	0.0899	0.7583	0.7097
AYK	32.45	0.0970	0.8494	0.8356
LAY	29.11	0.0972	0.8156	0.7955
ND	24.92	0.0876	0.7563	0.7037
NOW	30.86	0.0995	0.8455	0.8185
UWN	32.20	0.0953	0.8417	0.8188
YAE	29.78	0.0976	0.8383	0.8094
YUW	27.18	0.0960	0.8223	0.7812

Table 2. . EER (%), minDCF, Cllr and minCllr for the 10 best performing diphone units in the NIST SRE 2006 English-only male 1side-1side task.

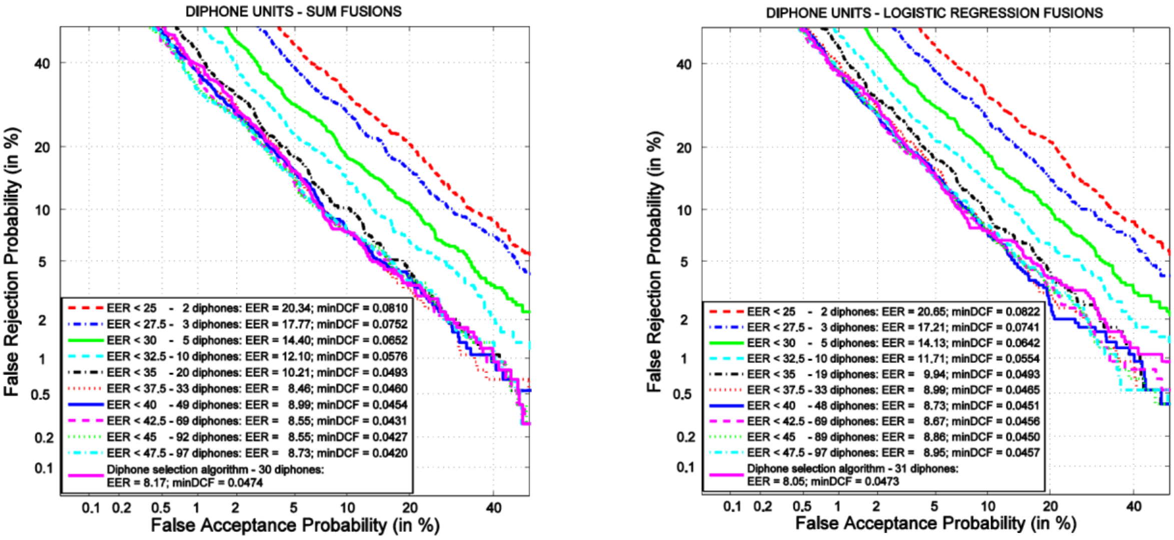

Fig. 3. DET curves for sum and log. reg. fused systems in the NIST SRE 2006 English-only male 1side-1side task for different diphone selection schemes.

5.4 Inter-unit combined system performance

In the previous paragraphs we have seen how well combine different units from each type (i.e., different phones between them and different diphones between them), but it is also interesting to see how can be combined units from different types between them. For this purpose, same fusion techniques and combination schemes have been used putting together both phones and diphones, yielding results show in Figure 4.

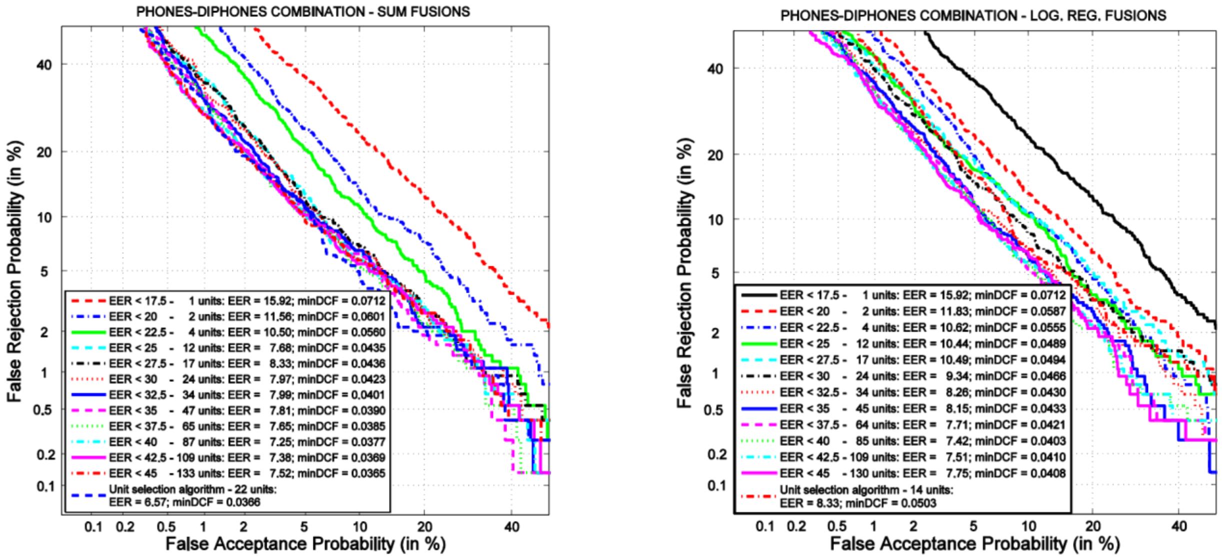

Fig. 4. DET curves for sum and log. reg. fused systems in the NIST SRE 2006 English-only male 1side-1side task for different phone-diphone selection schemes.

It can be seen that better results can be achieve by combining phones and diphones units than working in a intra-unit manner, taking advantage of different linguistic levels. This way, it is possible to achieve improvements around 35% in terms of EER over the reference system, as it can be seen in Table 3.

6 Summary and conclusions

In this paper we have presented an analysis of the contributions of individual linguistic units to automatic speaker recognition by means of their cepstral trajectories, showing that some of them can be used to obtain informative likelihood ratios very useful in forensic applications, with the advantage of being a completely automatic system and using parameters similar to those used by linguists or phoneticians. This way it is possible to deal with uncontrolled scenarios where only some short segments are available to be compared, making it possible to infer a conclusion about the speaker identity in the speech sample. This procedure cannot be done by the usual automatic speaker recognition systems because they use all available speech data as a whole, and usually they are tuned to work with fixed-length training and testing segments. Furthermore, when more testing data is available, individual units can be combined to improve the discrimination capabilities of the resulting system, having shown that these combinations, both at intra- and inter-unit levels, can outperform the results obtained with the same system framework based on MFCC features.

System	# fused units	EER (%)	minDCF
Reference	-	10.26	0.0457

Phones – best fused system (sum)	17	7.11	0.0420
Diphones – best fused system (log. reg.)	31	8.05	0.0473
Phones+diphones – best fused system (sum)	22	6.57	0.0366

Table 3. Performance comparison between the reference system and unit-based fused systems in the NIST SRE 2006 English-only male 1side-1side task

7 References

1. Brummer, N. et al., “Application-independent evaluation of speaker detection”, Comp. Speech Lang., (20) 230-275, 2006.
2. Dehak, N., et al., “Front-End Factor Analysis for Speaker Verification”, IEEE Trans. on Audio, Speech and Lang. Proc., 19(4), 788-798, May 2011.
3. Castro, A. d., Ramos, D., and Gonzalez-Rodriguez, J., “Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking”, in Proceedings of Interspeech 2009, pp. 2343-2346, September 2009.
4. Ferrer, L., “Statistical modeling of heterogeneous features for speech processing tasks”, Ph.D. dissertation, Stanford Univ., 2009 (http://www.speech.sri.com/people/lferrer/thesis.html)
5. Franco-Pedroso, J., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J., and Ramos, D. “Fine-grained automatic speaker recognition using cepstral trajectories in pone units”. Proceedings of IAFPA 2012, Santander, Spain.
6. Kajarekar, S. et al., “The SRI NIST 2008 Speaker Recognition Evaluation System”, Proc. IEEE ICASSP’09, pp. 4205-4209, Taipei, 2009.
7. Kenny, P. et al., “A Study of Inter-speaker Variability in Speaker Verification”, IEEE Trans. on Audio, Speech and Lang. Proc., 16(5):980-988, 2008.
8. Kenny, P., “Bayesian speaker verification with heavy tailed priors”, Keynote presentation at Odyssey 2010, Brno, 2010.
9. Kinnunen, T., and Li, H., “An overview of text-independent speaker recognition: from features to supervectors”, Speech Communication, vol. 52, pp. 12-40, 2010.
10. NIST SRE 2006 Evaluation Plan: http://www.itl.nist.gov/iad/mig/tests/sre/2006/sre06_evalplan-v9.pdf
11. Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., “Speaker verification using adapted gaussian mixture models”, Digital Signal Processing 10, pp. 19-41, 2000.
12. Shriberg, E., “Modeling prosodic feature sequences for speaker recognition”, Speech Communication, 46 (3-4), July 2005, pp. 455-472, Jan. 2005.
13. Wikipedia contributors. “Arpabet”. *Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/wiki/Arpabet (19 July, 2012.)