Speaker recognition using temporal trajectories in linguistic units: the case of formant and formant-bandwidth contours

Joaquin Gonzalez-Rodriguez [1,2]

[1] International Computer Science Institute, Berkeley, CA, USA
[2] ATVS, Universidad Autonoma de Madrid, Spain
joaquin@icsi.berkeley.edu / joaquin.gonzalez@uam.es

Abstract

We describe a new approach to automatic speaker recognition based in explicit modeling of temporal contours in linguistic units (TCLU). Inspired in successful work in forensic speaker identification, we extend the approach to design a fully automatic system, illustrated here with formant and bandwidth trajectory features, with a high potential for combination with acoustic-spectral systems. Using SRI's Decipher phone, word and syllabic labels, we have tested up to 468 unit-based subsystems from 6 groups of lexically-determined units, namely phones, diphones, triphones, center phone in triphones, syllables and words, subsystems being combined at the score level. Evaluating with the NIST SRE04 English-only 1s1s task, the hierarchical fusion of those groups result in an EER of 4.20% (minDCF=0.018), a remarkable result given the limited performance of automatic formant estimation in conversational telephone speech. Combining extremely well with a Joint Factor Analysis cepstral system (from JFA EER of 4.25% to 2.47%, minDCF from 0.020 to 0.012), future extensions to more robust prosodic and/or spectral features are likely to further improve this approach.

Index Terms: speaker recognition, linguistic units, temporal trajectories, formants, bandwidths.

1. Introduction

Speaker formant dynamics have been selected for study as initial feature for several reasons. Formant analysis has a long tradition in forensic phonetics, and they are features that linguists and phoneticians are comfortable with when defending them in court. Formant frequencies and their dynamics have shown strong individualization potential [Nol83][Ros02][McDou06], and different researchers (mostly linguists and phoneticians) [Mor09] [Zha08] [Kin01] [Ros10] [Cas09] have shown how to report Likelihood Ratios from formant trajectories, complying with most of the requisites of modern forensic science. In the above approach to LR reporting through formant trajectories, as formant frequencies are manually extracted and/or supervised, and the speech data in use have been recorded in controlled conditions, their approach shows two main limitations:
Very limited size of the data can be processed, as huge amounts of human work is needed, e.g., they record specific reference populations for every forensic case trying to match specific linguistic events and acoustic conditions present in the questioned speech. As a result, only very limited controlled data per target and reference speaker is available, only small set of linguistic features (phonemes, diphthongs) are evaluated, and just tens of speakers are usually involved in the experiments/reports.
The human expert has a strong influence in the selection of segments and assignment and/or correction of formant frequencies and trajectories, which severely limits the desired forensic transparency, testability and repeatability of the approach.
This project is the first attempt, to the author knowledge, to recognize speakers from formant trajectories in a fully automatic (testable) way with actual conversational speech (NIST SRE data), involving hundreds of male and female speakers with SRE-size number of trials (in the order of tens of thousands per eval), and with the possibility to use most or all of the available speech in each utterance, as any phone/diphone (or other units) can be included in the analysis.
The objective of this research is to develop a cepstrum-orthogonal automatic system (as e.g. prosodic ones), with high potential of fusion with state-of-the-art cepstral systems because of the different nature and time span of the features under analysis (formant and bandwidths trajectories initially).

2. Selection of lexically-determined units

Looking for multiple separate contributions to the speaker identity in a speech file, linguistic units are the natural and straightforward group of segments to work with. Using SRI's Decipher labels, six groups of units have been explored, showing each of them different characteristics in term of speaker identification from their formant and bandwidth trajectories specificities:

- *Phones*: showing the biggest frequencies of occurrence (from 20 to over 100 per conversation) among the 6 groups, they are also highly dependent on their contexts. Also within-phone formant and bandwidth trajectories show limited excursions, so less speaker-dependent clues are to be found, apart from target formant frequencies and bandwidths. Along the paper they will be noted with their Decipher symbol, e.g., AE, AY, DH, N, OW, T, etc.
- *Diphones*: they show on average richer contours than phones but poorer than triphones. Occurrence frequencies range from units (3-4) to some tens (20-25). The good balance between occurrences and rich trajectories will give us the best performing individual units from all six groups. Notation: DH+AE, AE+T, AY+N, N+OW, etc.
- *Triphones*: they show the richer contours on average but their frequency of occurrence drops dramatically (just 20 of them occur on average more than three times per conversation). Additionally, the sometimes high number of articulation targets makes the resulting contour not adequate to be coded with just 5-DCT coefficients. hiNotation: DH+AE+T, AY+N+OW, etc.
- *Center phone in triphones*: modeling just the central phone, higher control of the context makes them attractive, but with the same low frequency as triphones and shorter units being modeled (center phone instead of

triphones) only some of them will be finally useful. Decision-tree-based clustering of triphones should be useful in a near future with this set of units, just like triphone modeling in speech recognition, learning from the data which triphones are similar enough to group together in terms of the formant and bandwidth contours in the attack and release of the center phone, thereby dealing with the data sparseness of those units. Notation: dh+AE+t, ay+N+ow, etc.

- *Syllables*: they show high frequency of appearance, sharing some of the best performing units with phones, diphones and triphones but having less contextual variation (e.g., a diphone is extracted in any context, even being the last and beginning phone in consecutive words, while a diphone syllable is only extracted when detected as such). Their high linguistic consistency, similar to that of words below, combined with their good frequency of occurrence will result in the best performing group of units. Notation: ay, n+ow, dh+ae+t, etc.
- *Words*: only a few of them are frequent enough to perform well (function words as “but”, backchannels as "yeah", fillers like "uh", discourse markers like "so", etc.) but they can be idiosyncratic for speakers. They are also often surrounded on one or both sides by a pause, which helps reduce contextual variation. Longer words may be more stable inside the word, due to less contextual variation away from the word edges, especially on a stressed syllable, but such words are also much less frequent. Notation: I, KNOW, THAT, etc.

Units with enough frequency of occurrence have been extracted from the SRE04 labels and tested in terms of EER, DCF, Cllr and minCllr, selecting those with comparatively good speaker identification capacities.

3. System description

Authors should observe the following rules for page layout. A highly recommended way to meet these requirements is to use a given template (OpenOffice, Word® or LaTeX) and check details against the corresponding example file.

3.1. Segmentation with SRI’s Decipher & Syllabifier

In order to segment the different units in use in our system, we use the phonetic and word transcription labels produced by SRI´s Decipher conversational telephone speech recognition system [Sto08]. The Word Error Rate (WER) of native and nonnative speakers on transcribed parts of the Mixer corpus, equivalent to the NIST SRE04 data used along this paper, was 23.0% and 36.1% respectively. However, in an informal evaluation performed by the author with 20 randomly selected files from native speakers in the SRE04 evaluation set, the perceived quality of the transcription is much better, in any case far from the one error every four words suggested by the above WERs.

Recently, the SRI syllabifier was improved (*with XX YY ZZ … –optional text here-*) for the development of the constrained cepstral system described in [BocXY]. This improved syllabifier was used to produce new syllable labels for the SRE04 data in use in this work.

3.2. Formant and bandwidth extraction

Automatic formant and bandwidth estimation from conversational telephone speech is a really challenging task, critical to the performance of the proposed system. The formant and formant-bandwidths estimator included in Wavesurfer [Sjo00] is the tracker in use in all the experiments described in this paper. It estimates speech formant trajectories through dynamic programming, used to optimize trajectory estimates by imposing frequency continuity constraints. The formant frequencies are selected from candidates proposed by solving for the roots of the linear predictor polynomial computed periodically. The local costs of all possible mappings of the complex roots to formant frequencies are computed at each frame based on the frequencies and bandwidths of the component formants for each mapping. The cost of connecting each of these mappings with each of the mappings in the previous frame is then minimized using a modified Viterbi algorithm.

Figure 1 show a sample performance of the formant trackers, where as can be seen, the performance is largely dependent of the unit and the formant under analysis. Moreover, the precision of the speech transcription alignment shows a tolerance of some 10ms frames, irrelevant for the purpose of transcription, but this tolerance introduces significant errors in the extracted contours associated with each label, especially in the transitions to/from plosive sounds, or between speech and non-speech segments.

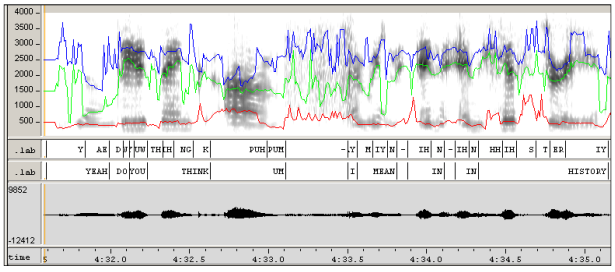

Figure 1: *formant contours (F1, F2, F3) aligned with SRI Decipher´s phone and word labels in a sample utterance from the SRE04 1side1side task (formants extracted and plotted with Wavesurfer [Sjo00]).*

On the other hand, figure 2 shows sample temporal contours of formant bandwidths, which have been shown beneficial for the purpose of speaker recognition and with speaker discrimination capabilities in themselves, as will be shown in the Results section.

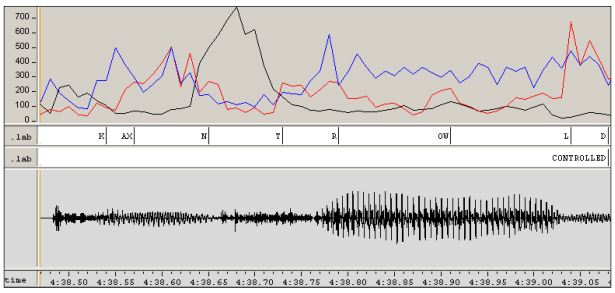

Figure 2: *Formant-bandwidth frequency (Hz) contours of first three formants (B1: Black, B2: Red, B3: Blue) aligned with SRI phone and word labels in a sample speech utterance from SRE04 1s1s task (extracted and plotted with Wavesurfer [Sjo00]).*

3.3. Unit parameterization

Once a speech unit is specified (e.g., diphthong, diphone, syllable …), formants and bandwidth contours are extracted, consistent of variable length sequences of 6 coefficients (3

formants and 3 bandwidths) every 10 milisecond frame. Following previous studies [Morrison][Castro], they are duration equalized to 250 ms, as is has been shown better performance with this equalization. However, as different speakers produce different duration patterns, the removal of this equalization step and the use of duration as an extra feature coefficient should be subject of future research.

Those segment trajectories are usually parameterized in previous works [REFS] through polynomial fitting or discrete cosine transform (DCT) coding. DCT was selected because of the pseudo-orthogonal properties of DCT coefficients for further statistical modeling. In previous works, order 3 DCT is commonly used, but because of the wide spectrum of units under evaluation, some peculiarities from the contours are lost with this 3-DCT configuration, so all our contours are coded with 5-DCT (c0, c1 .. c4) coefficients. It is of interest that the 5-DCT coding, apart from extracting the speaker contours, also partially compensates the noisy trajectory estimates from the formant tracker.

3.4. MVLR modeling and scoring: MVN & MVK

In order to model and score individual units of information, direct MultiVariate Likelihood Ratio (MVLR) generative methods known as MVN (MV Normal) and MVK (MV Kernel densities) in [Ait05] are used. Our MVN assumes a multivariate full covariance Gaussian model for both target and background models, with within- and between-speaker covariance matrices estimated from the background data, while in MVK, using the wording in [Mor11], "… *the between speaker distribution is modeled via the summation of a set of equally-weighted kernels with one kernel per speaker centered on the mean vector of the measurements from that speaker. Each kernel is a Gaussian whose covariance matrix is a scaled version of the pooled within-group covariance matrix. The scaling, and hence the degree of kernel smoothing, is determined by a function of the number of groups in the background database*".

While recent results with a semiautomatic approach with five diphthongs in a read-speech high-quality-microphone-recording 27-male-speakers dataset suggests comparable performance in terms of C_{llr} of a GMM-UBM approach relative to MVK (table 1 in [Mor11]), the MVLR techniques are preferred here for their inherent ability to produce calibrated likelihood ratios without the need for explicit data-based calibration procedures, one of the main reasons for MVLR being widely used in different forensic disciplines.

4. Datasets and experimental setup

As the definition and extraction of the different groups of units rely on word recognition, being thus language dependent, we use the English-only subset of the NIST SRE04 1side1side task, which comprises both native and nonnative speakers across 9,655 same-sex different-telephone-number trials from 208 speakers (123 female and 85 male) in 1,384 5-minute conversation sides (802 female and 582 male), each conversation side consisting of the last five minutes of a six-minute conversation, eliminating the less-topical introductory dialogue. All these conversational speech data, totalizing circa 114 hours of conversations (about 57 hours of net speech to be processed assuming 50% of speech per conversation side) was collected in the Mixer Project using the Linguistic Data Consortium's "Fishboard" platform. The cost function in use is the "old" well-known CDET as defined in the SRE04 eval plan [ref. NIST04].

The experimental setup has been designed such that for every two speakers involved in a trial (or one if it is the case of a target trial), reference background data for MVN/MVK and jackknife calibration training scores are extracted from all the remaining speakers/trials, guaranteeing that no speech or trials of the test speakers are known in any sense to the system, even from conversations different to the one involved in the trial.

5. Results

5.1. Features and MVLR method selection

Due to the different nature of each unit under analysis, not every feature of interest (three first formants and bandwidths) is properly extracted/tracked for all of them. Table 1 shows an exploratory analysis for 4 different diphthongs, where the best features in EER terms are different for different units. Especially remarkable, bandwidths themselves show speaker discrimination abilities, having all three configurations (B123, B12 and B23) minC_{llr} values smaller than one. However, for the sake of consistency, we have selected for the rest of the paper a constant 3 formants 3 bandwidths (FB123) MVK configuration for every unit. Finding the unit-dependent best configuration is an open door to further future improvements of the system.

	AY		EY		OW		AW	
Feat	**N**	**K**	**N**	**K**	**N**	**K**	**N**	**K**
F123	27,6	29,2	34,6	35,1	31,9	**29,1**	**36,3**	27,2
F12	28,1	29,7	33,0	34,3	**30,8**	29,5	36,4	32,1
F23	35,0	33,2	37,0	37,4	36,6	32,0	38,2	33,2
FB123	26,9	***25,6***	**31,4**	***30,6***	32,9	*30,3*	37,2	***23,5***
FB12	**26,7**	27,9	32,1	34,4	31,6	31,4	38,5	30,0
FB23	32,0	30,0	34,9	34,2	38,1	33,9	38,9	26,5
B123	34,6	33,4	36,1	37,4	39,9	36,5	42,8	36,9
B12	34,4	34,6	37,5	40,7	38,5	38,0	44,5	41,4
B23	37,8	37,2	38,1	39,8	42,0	39,2	45,5	40,7

Table 1: *EERs (%) of different combination of features (F: formants, B: bandwidths) with two multivariate LR methods (N: MVN, K: MVK) for 4 diphtongs in SRE04 english 1s1s male trials.*

5.2. Performance of individual linguistic units

A total of 468 units, consisting in X phones, Y diphones, Z triphones, A center phone in triphone, B words and C syllables have been explored. Figure 3 shows a strong correlation of EER per unit with low frequencies of occurrence. However, for average frequencies over a certain threshold (about 3 occurrences per conversation, log10(3)~0.5, the left column of data in the graph), the performance is solely dependent on the acoustic characteristics of every unit and not its frequency.

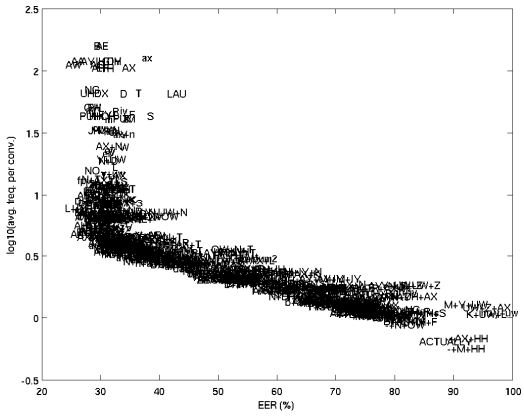

Figure 3: *Scatter plot of EERs(%) versus average frequency (in log10 units) for the 468 explored units in the SRE04 English-only 1s1s task.*

Table 2 shows that among the 40 best individually performing units we find mostly diphones (19) but all six groups of units are represented.

UNIT	EER(%)	Freq.	UNIT	EER(%)	Freq.
L+IY	24,02	7,83	PUH	26,45	44,07
AW	24,10	113,24	UH	26,53	66,93
AA+T	24,96	4,85	dh+ae+t	26,64	8,39
AA	25,06	121,4	AY	26,66	121,4
l+iy	25,18	7,77	t+uw	26,75	5,03
AH+M	25,52	5,67	N+T	26,87	13,67
DH+AE	25,63	9,02	IH+N	26,87	5,53
EH+N	25,63	7,26	TH+IH+NG	26,87	6,48
AY+M	25,74	6,68	AE+N+D	27,06	7,39
BUT	25,95	5,05	L+AY	27,09	10,85
IH+NG	25,96	6,91	OW	27,21	51,56
T+UW	25,96	5,04	NG	27,32	71,46
DH+EH	26,04	6,70	N+OW	27,32	11,08
AX+N+T	26,07	4,58	AX+NG	27,32	9,28
AE+T	26,08	10,09	B+AX	27,32	6,06
fpv	26,10	13,40	NO	27,32	16,06
AO+R	26,19	7,28	NOW	27,42	10,32
DH+AE+T	26,21	8,71	n+ow	27,50	10,97
AH+N	26,31	5,19	dh+AE+t	27,54	8,71
AH+T	26,44	7,22	ay	27,54	49,11

Table 2: *EERs (%) and average frequency of occurrence of the 40 best performing units in EER, in the SRE04 English 1side1side task.*

5.3. Results with groups of linguistic units

In table 3 we combine subsystems within a group of units based in the identification performance (in EER terms) of the units. It is remarkable that all six groups perform reasonably well in EER in this SRE04 task, which is extensible to the calibration loss (C_{llr} – $minC_{llr}$), with the exception of *diphones* whose calibration loss degrades for higher number of units.

EER<	# units	EER (%)	100xDCF	Cllr	minCllr
29%	14 phones	14,89%	6,3274	0,5431	0,4928
	38 diphones	11,34%	4,2321	0,4301	0,3600
	5 triphones	17,86%	6,1236	0,5927	0,5354
	3 ph_triph	22,93%	7,735	0,7081	0,6554
	7 words	19,22%	6,5718	0,6304	0,5751
	15 syllables	13,52%	5,5744	0,4759	0,4334
31%	24 phones	13,7%	6,4299	0,5104	0,4636
	70 diphones	8,43%	3,6833	0,4851	0,2972
	13 triphones	14,01%	5,1446	0,4978	0,4426
	4 ph_triph	19,78%	7,0248	0,6465	0,5938
	13 words	14,62%	5,6057	0,5019	0,4603
	31 syllables	10,78%	5,0679	0,4405	0,3753
36%	41 phones	8,978%	4,6059	0,3723	0,3238
	91 diphones	7,358%	3,1751	0,7345	0,2685
	24 triphones	11,91%	4,6837	0,4429	0,3893
	9 ph_triph	16,86%	5,809	0,5629	0,5070
	16 words	13,93%	5,1562	0,4830	0,4404
	49 syllables	**6,913%**	**3,0648**	**0,3232**	**0,2169**

Table 3: *EERs (%), DCF, C_{llr} and $minC_{llr}$ for different within-group combinations of unit subsystems in the SRE04 English 1side1side task.*

Even though the *diphone* group seems a good competitor to the best-performing *syllable*-group (*diphones* "winning" in two of the above conditions in EER terms, and close to the *syllables* in the third), this is at the expense of the higher calibration loss among all the groups in all the cases, what strongly correlates in our experiments with the need for a larger number of units to obtain good discrimination results.

5.4. Fusion of linguistic units

Two types of fusion experiments across groups of units have been performed. Table 4 shows the first of them, where with the same selection criteria as in table 3, units are selected independently of the group they belong to.

EER<	#units	M/F	EER (%)	100xDCF	Cllr	minCllr
29%	82	M	9.99%	4.4465	0.4356	0.3256
		F	9.14%	3.6889	0.4062	0.2909
		M+F	9.43%	4.0884	0.4189	0.3121
31%	155	M	7.94%	3.2342	1.592	0.2632
		F	5.86%	2.3524	1.653	0.2082
		M+F	6.61%	2.7920	1.623	0.2378
36%	230	M	7.91%	3.9451	6.565	0.2779
		F	4.86%	1.6963	7.243	0.1463
		M+F	**6.39%**	**2.7048**	**7.019**	**0.2146**

Table 4: *EERs (%), minDCF, C_{llr} and $minC_{llr}$ in the SRE04 English 1side1side task (M: male, F: female).*

Due to the different nature of the units to be combined, even while results in terms of EER are good, the calibration gets horrible, meaning that strongly misleading Likelihood Ratios are provided from time to time.

The second set of across-group experiments is summarized in table 5, where the best groups of units are hierarchically fused (group by group) in a sequence order given by group performance.

EER<	# units	EER	100xDCF	Cllr	$minC_{llr}$
29%	(1) = 15syll+38diph	4.72	1.9279	0.2140	0.1654
	(2) = (1) + 14ph	**4.20**	1.8203	**0.2023**	0.1606
	(3) = (2) + 5triph	4.12	**1.7771**	0.2119	0.1617
	(4) = (3) + 7words	3.88	1.9397	0.2097	**0.1590**
	(5) = (4) + 3phtriph	3.88	2.0591	0.2117	0.1618
31%	(1) = 31syll+70diph	5.70	2.4403	0.2559	0.2090
	(2) = (1) + 24ph	5.51	2.2105	0.2304	0.1920
	(3) = (2) + 13triph	5.41	2.3052	0.2308	0.1932
	(4) = (3) + 13words	5.54	2.2771	0.2353	0.1972
	(5) = (4) + 4phtriph	5.64	2.2751	0.2401	0.1997
36%	(1) = 49syll+91diph	5.04	2.1955	0.2327	0.1895
	(2) = (1) + 41ph	4.81	**2.0272**	**0.2147**	**0.1792**
	(3) = (2) + 24triph	4.93	2.0497	0.2200	0.1806
	(4) = (3) + 16words	**4.60**	2.1065	0.2208	0.1799
	(5) = (4) + 9phtriph	4.71	2.1627	0.2242	0.1832

Table 5: *EERs (%) and average frequency of occurrence of the 40 best performing units in EER, in the SRE04 English 1side1side task.*

Interestingly, while excellent results are obtained with large number of units from all groups (e.g., 221 units for EER=4.60%), the best results (EER=4.20%, DCF=0.0178, Cllr=0.2023) are obtained fusing a limited number of units, ranging from 67 (15syll+38diph+14ph) to 79 (idem+5triph+ 7words). Moreover, the calibration loss and actual C_{llr} are good in all the above cases, whatever configuration is chosen,

which means that all of them are always providing informative Likelihood Ratios.

5.5. Fusion with a JFA cepstral system

Among the many desirable properties of higher level systems [Lizlibro][Lizforensic], nice fusion complementarities with cepstral systems is one of the most prominent, exploiting identification clues totally unexplored by the frame-by-frame underlying scoring (real or virtual) of all of them. Table 6 shows the performance of one of our best TCLU (Temporal Contours of Linguistic Units) systems compared to a standard cepstral GMM-MAP (raw scores, no ZTnorm) and a 50 eigenchannels Joint Factor Analysis systems.

System	M/F	EER (%)	100xDCF
(1) Linguistic contour (TCLU)	M	5.539%	2.2771
	F	3.879%	1.9397
	M+F	**4.60%**	**2.1065**
(2) GMM-MAP	M	14.09%	6.0675
	F	14.07%	5.7737
	M+F	14.01%	5.9585
(3) JFA u50	M	4.669%	2.1061
	F	3.98%	1.542
	M+F	**4.255%**	**1.9953**
Sum fusion: (1) + (3)	M	2.743%	1.4271
	F	2.191%	0.8833
	M+F	**2.475%**	**1.1161**

Table 6: *EERs (%) and average frequency of occurrence of the 40 best performing units in EER, in the SRE04 English 1side1side task.*

Not only the results of the cepstral and linguistic contours systems are fully comparable (see figure 4), but a simple non-trained sum fusion of both system is outstanding, showing once more the complementarities of low and higher level approaches.

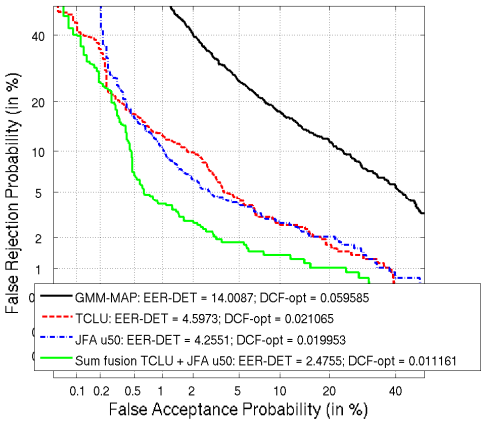

Figure 4: *DET plots of EERs(%) versus average frequency (in log10 units) for the 468 explored units in the SRE04 English-only 1s1s task.*

6. Discussion & conclusions

The identification performance (EER) of the different units, with a system based on modeling the trajectories of formants and bandwidths within the unit, depend on:
- how large formant excursions are within the unit, large excursions favouring speaker peculiarities
- the frequency of appearance of the unit in the conversation, in order to have reliable speaker models and scoring
- the number of articulation targets (~ #phones) in the unit: modeling each trajectory with 5 DCT coefficients, complex units (e.g. long words) can have complex contours that are not properly covered with just 5 coefficients, and short units (phones) become highly dependent on the context.
- the consistency of the unit both during the conversation and intersession: syllables and words, having a higher linguistic "structure" (?), are expected to be more consistent than shorter units in different contexts

Combining all those facts, and this is just my guess (inspired from results, see below):
- phones perform well because of very high frequency of appearance, but high context dependency induces variability
- diphones (some of them) are the single best performing units, but as a group perform slightly worse than syllables
- triphones are very promising but their frequency of appearance drops dramatically
- center phone in triphones are even more promising, but with the same low freq as triphones and shorter units being modeled (phones instead of triphones) only some of them are useful
- words: only a few of them are frequent and short enough to perform well
- syllables: preserve some of the good properties of words, show high frequency of appearance and have enough formant excursion within because of the number of articulation targets.

7. Acknowledgements

The author thanks Prof. Nelson Morgan and the speech group at ICSI for hosting and supporting this work during the 2010-2011 academic year. Special thanks to Liz Shriberg and Andreas Stolcke for suggestions and encouragement, to Luciana Ferrer and SRI for providing the Decipher labels, to Howard Lei and Lara Stoll for helping with system development and data management, and to all ATVS members for remote support. Thanks to Niko Brummer for the FoCal toolkit and Geoff Morrison for the MVK implementation. This research has been supported by Ministerio de Educacion research stay grant PR-2010-123, MICINN project TEC09-14179, and the research activities in Catedra UAM-Telefonica.

8. References

[1] [Ait05] C. G. G. Aitken and D. Lucy, ”Evaluation of trace evidence in the form of multivariate data”, Applied Statistics 53, pp. 109-122, with corrigendum pp. 665-666, 2005.

[2] [Cas09] A. de Castro, D. Ramos and J. Gonzalez-Rodriguez, “Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking”, Proc. Interspeech 2009, Brighton, UK, 2009, pp. 2343-2346.

[3] [Kin01] Kinoshita, Y. “Testing Realistic Forensic Speaker Identification In Japanese: A Likelihood Ratio Based Approach Using Formants”. Linguistics. (2001) Canberra, The Australian National University.

[4] [McDou06] K. McDougall, “Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies”, Int. Jour. on Speech Language and the Law 13(1), pp. 89-126, 2006.

[5] [McDou07] K. McDougall and F. Nolan, “Discrimination of Speakers using the Formant Dynamics of /u:/ in British English”, Proc. of Int. Conf. on Phonetic Science (ICPhS), Saarbrucken, August 2007, pp. 1825-1828.

[6] [Mor09] Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories. Morrison, G. S. (2009). Journal of the Acoustical Society of America, 125, 2387–2397.

[7] [Nol83] F. Nolan, “The Phonetic bases of speaker recognition”, Cambridge University Press, Cambridge (UK), 1983.
[8] [Ros02] P. Rose, Forensic Speaker Identification, Taylor & Francis, 2002.
[9] [Ros10] P. Rose, “The effect of correlation on strength of evidence estimates in Forensic voice comparison: uni- and multivariate Likelihood Ratio-based discrimination with Australian English vowel acoustics”, Int. Journal of Biometrics, Vol. 2, No. 4, pp. 316-329, 2010.
[10] [Rud07] D. Rudoy, D.N. Spendley and P.J. Wolfe, “Conditionally linear gaussian models for estimating vocal tract resonances”, Proc. Interspeech 2007, Antwerp, Belgium, 2007, pp. 526-529.
[11] [Sjo00] K. Sjolander and J. Beskow, “Wavesurfer – an open source speech tool”, Proc. ICSLP 2000, Bijing, China, 2000.
[12] [Zha08] Zhang, C., Morrison, G. S., & Rose, P., “Forensic speaker recognition of Chinese /i/ and /y/ using likelihood ratios”. (2008). Proceedings of Interspeech 2008 (pp. 1937–1940). International Speech Communication Association.
[13] [Liz07] Shriberg, E., “Higher-level features in speaker recognition”, in Speaker Classification I: Fundamentals, Features and Methods, C. müller, Ed., number 4343 in Lecture notes in Artificial Intelligence, pp. 241-259, Springer, 007.
[14] [Sto08] Stolcke, A. et al., “The SRI March 2000 Hub-5 conversational speech transcription system”, in Proc. NIST Speech Transcription Workshop, College Park, MD., 2008.
[15] [Boc09] Bocklet, T. and Shriberg, E., “Speaker recognition using syllable-based constraints for cepstral frame selection”, Proc. ICASSP’09, YYYY ZZZZ.
[16] [Mor11] Morrison, G., “ A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: MVKD versus GMM-UBM”, Speech Communication, 53: 242-256, 2011.
[17] [Shr08] Shriberg, E. and Stolcke, A., "The case for automatic higher-level features in forensic speaker recognition", Proc. *INTERSPEECH’08*, 1509-1512, 2008.
[18] [SRE04] NIST, http://www.itl.nist.gov/iad/mig/tests/sre/2004/SRE-04_evalplan-v1a.pdf
[19] [SRI09] S. S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, & T. Bocklet (2009), The SRI NIST 2008 Speaker Recognition Evaluation System Proc. IEEE ICASSP/, pp. 4205-4209, Taipei.