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

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, with a high potential for combination with 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 NIST SRE04 English-only 1s1s, their hierarchical fusion gives an EER of 4.20% (minDCF=0.018) from automatic formant tracking of conversational telephone speech. Combining extremely well with a Joint Factor Analysis system (from JFA EER of 4.25% to 2.47%, minDCF from 0.020 to 0.012), extensions as more robust prosodic or spectral features are likely to further improve this approach.

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

1. Introduction

The use of higher level features for speaker recognition [1] have shown multiple desirable properties, ranging from discriminative power and potential for combination with short term spectral systems, to interpretability and acceptance, which make them of general interest but especially suitable for forensics [2]. Among the variety of features available [1], largely unexplored speaker formant dynamics, especially from an automatic perspective, have been selected here for study. 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 [3][4], and different researchers, mostly linguists and phoneticians following the pioneering steps of Phil Rose [5][6][7][8][9][10], have shown how to report Likelihood Ratios from formant trajectories, complying with most of the requisites of modern forensic science [11][5]. However, as formant frequencies are manually extracted and/or supervised for every linguistic unit of interest, a very limited percentage of the available data can be processed, as huge amounts of human work is needed.

The objective of this work is to develop a cepstrum-orthogonal automatic system (as e.g. prosodic ones), with high potential of fusion with state-of-the-art spectral systems because of the different nature and time span of the features under analysis (formant and bandwidths trajectories initially), and with good properties in terms of calibration. This work is the first attempt, to the author knowledge, to recognize speakers from formant trajectories in a fully automatic way from actual conversational speech involving hundreds of male and female speakers, with large number of trials and with the possibility of using most or all of the available speech in each utterance, as any linguistic unit can be included in the analysis.

The remainder of the paper is organized as follows. In Section 2 we introduce the sets of units under analysis, while in section 3 we describe the different components of the proposed TCLU system. Sections 4 and 5 deal with the experimental part of the paper, showing results for a variety of conditions and combinations of the available units, to finally conclude in Section 6 summarizing the main contributions and identifying elements that could further improve the system.

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 formant-bandwidth trajectories specificities:

- *Phones*: showing the biggest frequencies of occurrence among the six groups (from 20 to over 100 per conversation), they are highly dependent on their contexts. Additionally, within-phone formant and bandwidth trajectories show limited excursions (except in the case of diphthongs).
- *Diphones*: they show on average richer contours than *phones* but poorer than *triphones*, presenting good enough occurrence frequencies (from units to 20-30).
- *Triphones*: they show the richer contours on average, but their frequency of occurrence drops dramatically. Additionally, some of them have high number of articulation targets resulting in complex contours to be modeled with a small number of parameters.
- *Center phone in triphones*: we extract the contours just from the central phone in a given triphone, limiting context variability. Being attractive, they share the same low frequency as *triphones* and the low number of articulation targets as *phones*.
- *Syllables*: they show both high frequency of appearance and rich contours, both of them desireable properties. They share some of the units with *phones*, *diphones* and *triphones* but as a group show less contextual variation.
- *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.

Copyright © 2011 ISCA

3. System description

3.1. Segmentation with SRI's Decipher & Syllabifier

In order to segment the different units, we use the phonetic and word transcription labels produced by SRI´s Decipher conversational telephone speech recognition system [12]. 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. The SRI syllabifier, developed for the constrained cepstral system in [13], was used with SRE04 data, where syllables are obtained using a maximum onset algorithm [14]. Given a stream of phones, onsets are detected by searching for matching phone sequences in a list of allowed English onsets obtained from an English lexicon. Syllables are completely determined by the set of onsets found.

3.2. Formant and bandwidth extraction

The formant and formant-bandwidths tracker included in Wavesurfer [15] is used in all the experiments in this paper. The formant frequencies are selected from candidates proposed by solving for the roots of the linear predictor polynomial computed periodically, estimating formant trajectories through dynamic programming, used to optimize trajectory estimates by imposing frequency continuity constraints. Unfortunately for our purposes, automatic formant contour estimation is a challenging task in conversational telephone speech, resulting usually in noisy contours. Moreover, the precision of the speech transcription, used to delimit unit boundaries, is far from perfect, adding contour artifacts in the edges of the units.

3.3. Unit parameterization

Formants and formant-bandwidth contours are duration equalized to 250 ms., following results in previous studies [6][9]. However, as different speakers produce different duration patterns, the removal of this equalization step and the use of duration information as an extra feature should be subject of future research. Those segment trajectories can be parameterized [6][9] through polynomial fitting or discrete cosine transform (DCT) coding, fifth order DCT being selected here because of the pseudo-orthogonal properties of DCT coefficients. Apart from extracting the speaker contours, removing the higher coefficients from DCT coding also helps filtering out noisy components in the original 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 [16] are used. Our MVN assumes a multivariate full covariance Gaussian model for both target and background models, with between-speaker covariance matrices estimated from background data, while MVK models the between speaker background data through a flat-weighted distribution of speaker centered Gaussian kernels.

While recent results suggests comparable performance in terms of C_{llr} of a GMM-UBM approach relative to MVK (table 1 in [17]), the MVN/MVK techniques are preferred here, instead of GMM-UBM, for their inherent ability to produce calibrated likelihood ratios without the need for explicit data-based calibration procedures, being MVN/MVK widely used as such 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). The Detection Cost Function (*DCF*) in use is the "old" C_{DET} as defined in the SRE04 eval plan, not the newest SRE10 one. All reported TCLU systems throughout this work elicit likelihood ratios, so C_{llr} and $minC_{llr}$ [18] (and its difference, calibration loss) have been largely used throughout the paper to evaluate the goodness of the different detectors.

The system under evaluation needs both reference background data for MVN/MVK [16], and subsystems trials scores for logistic regression fusion [18]. The experimental setup has been designed such that for every two speakers involved in a trial (or one if it is a target trial), both reference background data for MVN/MVK and jackknife logistic regression training scores are obtained from all the remaining speakers and trials, guaranteeing that no speech or trial scores involving any of the two speakers under evaluation are known in any sense to the system. All system components are highly efficient (even ASR [12]), enabling the whole system to be run in real time in a single multiprocessor computer.

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 show an exploratory analysis for 4 different diphthongs, where the best feature sets in terms of Equal Error Rate (EER) are different for different units.

	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 diphthongs in SRE04 English 1s1s male trials.*

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 all units under analysis, close enough to the best performing configuration in all the evaluated units. Finding the unit-dependent best configuration is an open door to further future improvements of the system.

Especially remarkable, bandwidths themselves show speaker discrimination abilities, having all three configurations (B123, B12 and B23) $minC_{llr}$ values smaller than one, therefore with potential to provide useful information to the user (if calibrated Likelihood Ratios are obtained).

5.2. Selection of individual linguistic units

A total of 468 units, consisting in 42 phones, 108 diphones, 129 triphones, 48 center phone in triphone, 30 words and 111 syllables have been explored. Units with enough frequency of occurrence are tested in terms of EER, DCF, C_{llr} and $minC_{llr}$, results which are summarized in figure 1, showing a strong correlation of unit's EER with low frequencies of occurrence. Once a minimum frequency of occurrence is guaranteed, performances are not so strongly dependent on frequencies. Equalizing the number of occurrences would be of interest to evaluate the frequency-independent goodness of the unit, but in actual conversations units and occurrences come together.

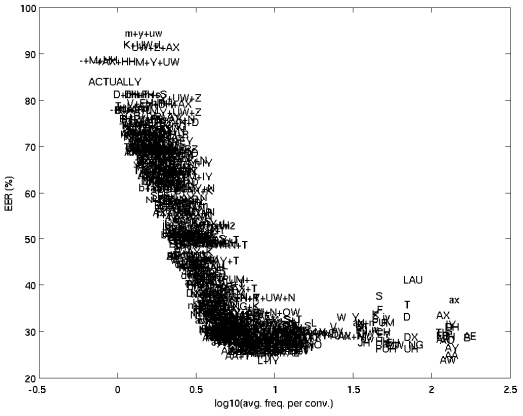

Figure 1: *Scatter plot of EERs (%) versus average frequency (in log10 units for visibility) for the 468 explored units in the SRE04 English-only 1s1s task.*

5.3. Results with groups of linguistic units

In table 2 we combine subsystems within a group of units based in the identification performance (in terms of EER) of the units. Being true that the selection criteria are designed in this work from the same data as later used as evaluation set, the underlying properties of the selected units are expected to be similar across different evaluation corpus.

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%	7,7350	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	13,70%	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	8,98%	4,6059	0,3723	0,3238
	91 diphones	7,36%	3,1751	*0,7345*	*0,2685*
	24 triphones	11,91%	4,6837	0,4429	0,3893
	9 ph_triph	16,86%	5,8090	0,5629	0,5070
	16 words	13,93%	5,1562	0,4830	0,4404
	49 syllables	**6,91%**	**3,0648**	**0,3232**	**0,2169**

Table 2: *EERs (%), DCF, C_{llr} and $minC_{llr}$ for different within-group combinations of unit subsystems in the SRE04 English 1side1side task. Unit selection criteria in shown in left column.*

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* where it degrades for higher number of units. Even though the *diphone* group seems in table 2 a good competitor to the best-performing *syllable*-group (*diphones* even being best group for the first two of the selection criteria in terms of EER, and close to the *syllables* EER performance in the third), this is achieved at the expense of a much higher calibration loss.

5.4. Fusion of linguistic units

Two types of fusion experiments across groups of units have been performed. Table 3 shows the first of them, where with the same selection criteria as in table 2, 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 3: *EERs (%), minDCF, C_{llr} and $minC_{llr}$ for across-group combinations of unit subsystems 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 seem good, the calibration gets increasingly worse (observe difference between C_{llr} and $minC_{llr}$ in table 3) for higher number of units to be combined, meaning that strongly misleading Likelihood Ratios are provided in the latter configuration.

The second set of across-group experiments is summarized in table 4, where the best groups of units are hierarchically fused (group by group), as described in section 4, in a sequence order given by group performance in terms of EER.

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

Interestingly, while excellent results are obtained with large number of units from all groups (e.g., 221 units give EER=4.60%), the best results (EER=4.20%, DCF=0.0178,

C_{llr}=0.2023) are obtained fusing a limited number of units, ranging from 67 (15syll+38diph+14ph) to 79 (idem67+ 5triph+ 7words). Moreover, the calibration loss and actual C_{llr} are good in all the cases in table 4, whatever configuration is chosen, which means that all those systems are providing informative meaningful Likelihood Ratios.

5.5. Fusion with a JFA cepstral system

Among the many desirable properties of higher level systems [1][2], nice fusion complementarities with short-term cepstral systems is one of the most prominents. Table 5 shows the performance of one of our best TCLU (Temporal Contours of Linguistic Units) systems compared to both a standard cepstral GMM-MAP and a 50 eigenchannels Joint Factor Analysis system (raw scores in both cases) in terms of EER and DCF.

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 (raw scores)	M	14.09%	6.0675
	F	14.07%	5.7737
	M+F	14.01%	5.9585
(3) JFA u50 (raw scores)	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 5: *EERs (%) and DCF for TCLU, cepstral systems and a combination of them in the SRE04 English 1side1side task (M: male, F: female)*

Not only the results of the cepstral and linguistic contours systems are comparable here, but fusion significantly improves the performance of each one of them individually, showing once more the complementarities of both approaches.

A simple non-trained sum fusion of JFA and TCLU systems was used here because of unavailability of "clean" (not used) scores to train logistic regression, sum fusion being done after monotonous transformations of both systems scores in order to have equivalent impostor score global distributions.

6. Summary and conclusions

We have described a higher level system that performs remarkably well in a common speaker recognition task, exploiting the combination of multiple pieces of information distributed among the linguistic units under analysis. We have shown that individual performance of units rely on a number of factors (contour excursions, frequency of units, context dependency, etc.), and that hierarchical combination of groups of units provide good discrimination and calibration in a variety of situations. The interest of the TCLU approach relies not just in its present performance, to be confirmed with further and more complex conditions showing, e.g, channel (mic vs. tel) or type of speech (conversational vs. interview) mismatch, but in the number of aspects identified for further future improvement, going from better formant estimation or the use of more robust prosodic or acoustic/spectral features, to more complex unit/group selection strategies or better grouping of units (e.g., decision-tree-based clustering of center phones in triphones can provide higher frequencies of occurrence of units with same good low context variability).

7. Acknowledgements

The author thanks Prof. Morgan and the speech group at ICSI for hosting and supporting this work during the 2010-2011 academic year. Special thanks to E. Shriberg and A. Stolcke for suggestions and encouragement, to L. Ferrer and SRI for providing the Decipher and syllable labels, to H. Lei and L. Stoll for help at ICSI, to D. Ramos and D. Torre for improving an early draft of this paper, and to all ATVS members for remote support. Thanks to Niko Brummer for the FoCal toolkit and Geoff Morrison for the MVK implementation. This work has been supported by MEC research stay grant PR-2010-123, MICINN project TEC09-14179, ForBayes project CCG10-UAM/TIC-5792 and Catedra UAM-Telefonica.

8. References

[1] 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, 2007.

[2] Shriberg, E. and Stolcke, A., "The case for automatic higher-level features in forensic speaker recognition", Proc. Interspeech'08, 1509-1512, Brisbane (AU), 2008.

[3] Nolan, F., "The Phonetic bases of speaker recognition", Cambridge University Press, Cambridge (UK), 1983.

[4] McDougall, K., "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] Rose, P., Forensic Speaker Identification, Taylor&Francis, 2002.

[6] Morrison, G. S., "Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories", J. of the Ac. Soc. of Am., 125, 2387–2397 (2009).

[7] Zhang, C., Morrison, G. S., & Rose, P., "Forensic speaker recognition of Chinese /i/ and /y/ using likelihood ratios". (2008). Proc. of Interspeech'08, 1937–1940, Brisbane (AU).

[8] Kinoshita, Y. "Testing Realistic Forensic Speaker Identification In Japanese: A Likelihood Ratio Based Approach Using Formants". Linguistics. (2001) Canberra, The Austr. Ntnl. Univ.

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

[10] Rose, P., "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.

[11] Gonzalez-Rodriguez, J. et al., "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition," IEEE Trans. on Audio, Speech and Language Processing, vol. 15, no. 7, pp. 2072–2084, 2007.

[12] Kajarekar, S. S. et al., "The SRI NIST 2008 Speaker Recognition Evaluation System", Proc. IEEE ICASSP'09, pp. 4205-4209, Taipei, 2009.

[13] Bocklet, T. and Shriberg, E., "Speaker recognition using syllable-based constraints for cepstral frame selection", Proc. ICASSP'09, Taipei, Taiwan, 2009.

[14] 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)

[15] K. Sjolander and J. Beskow, "Wavesurfer – an open source speech tool", Proc. ICSLP 2000, Bijing, China, 2000.

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

[17] Morrison, G. S., "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: MVKD versus GMM-UBM", Speech Comm., 53: 242-256, 2011.

[18] Brummer, N. et al., "Application-independent evaluation of speaker detection", Comp. Speech Lang., (20) 230-275, 2006.