Formant Trajectories in Linguistic Units for Text-Independent Speaker Recognition

Javier Franco-Pedroso, Fernando Espinoza-Cuadros and Joaquin Gonzalez-Rodriguez
ATVS - Biometric Recognition Group, EPS, Universidad Autonoma de Madrid
C/ Francisco Tomas y Valiente 11, 28049 Madrid, Spain
{javier.franco,joaquin.gonzalez}@uam.es, fernando.espinoza@estudiante.uam.es

Abstract

Inspired by successful work in forensic speaker identification, this work presents a higher level system for text-independent speaker recognition by means of the temporal trajectories of formant frequencies in linguistic units. Feature extraction from unit-dependent trajectories provides a very flexible system able to be applied in different scenarios. At a fine-grained level, it is possible to provide a calibrated likelihood ratio per linguistic unit under analysis (extremely useful in applications such as forensics), and at a coarse-grained level, the individual contributions of different units can be combined to obtain a more discriminative single system with high potential for combination with short term spectral systems. With development data being extracted from NIST SRE 2004 and 2005 datasets, this approach has been tested on NIST SRE 2006 1side-1side task, English-only male trials, consisting of 9,720 trials from 219 speakers. Remarkable results have been obtained for some single units from extremely short segments of speech, and the combination of several units leads to a relative improvement of 17.2% on EER when fusing with an i-vector system.

1. Introduction[1]

The use of higher level features for speaker recognition [14] has 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 specially suitable for forensics [15]. Formant analysis has a long tradition in forensic phonetics, and they are features that linguists and phoneticians are comfortable with when defending in court. Formant frequencies and their dynamics have shown strong individualization potential [9][11], and different researchers, mostly linguists and phoneticians following the pioneering steps of Phil Rose [3][10][13][18], have shown how to report likelihood ratios (LRs) from human-supervised formant trajectories, complying with the requisites of modern forensic science [13][6]. 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 amount of human work is needed.

In this work we present a fully automatic system based on formant trajectories extracted from linguistic units, 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, and having good properties in terms of calibration for forensic purposes. So, this paper present an alternative and complementary study to previous works [4][5] based on temporal trajectories of MFCCs.

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 different combinations, to finally conclude in Section 6 summarizing the main contributions of this work.

2. Feature extraction from linguistic units

Figure 1 shows the general idea of our feature extraction scheme. First, formant frequencies are computed by means of a formant tracker. Then, with the aid of an automatic speech recognition (ASR) system, linguistic units are bounded and formant frequency trajectories extracted. After a length normalization process, the temporal trajectory of the formant frequencies is encoded by means of a Discrete Cosine Transform (DCT) and the coefficients concatenated, yielding a constant-length feature vector per linguistic unit. Although several linguistic unit types can be analyzed, only phone and diphone units were used in this work.

2.1. Formant extraction

The Wavesurfer [16] formant tracker was used in order to extract the first three formant frequency values from the whole speech sample every 10 ms. Unfortunately for our purposes, automatic frequency contour estimation is a

[1] Supported by MEC grant PR-2010-123, MICINN project TEC09-14179, ForBayes project CCG10-UAM/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.

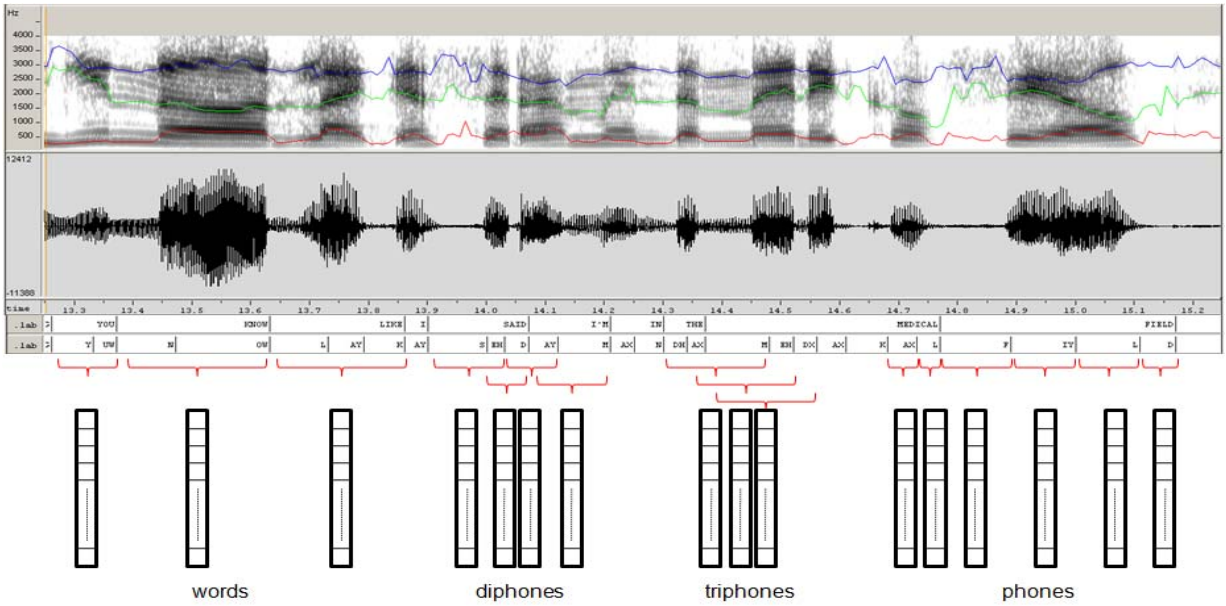

Figure 1: Constant-length feature vectors extraction from variable-length linguistic units.

challenging task in conversational telephone speech, resulting usually in noisy contours. Moreover, the precision of the speech transcriptions used to delimit unit boundaries is not perfect, adding contour artifacts in unit boundaries.

2.2. ASR region conditioning

Both phone and diphone units have been used for defining time intervals in order to extract the temporal contours of formant frequencies. For this purpose, the phonetic transcription labels produced by SRI's Decipher conversational telephone speech recognition system [7] were used. 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 units to be segmented. For this work, 41 phone units from an English lexicon were used, represented by the Arpabet phonetic transcription code [17]. Diphone units are defined by the combination of any two consecutive phone units, although only a subset of 98 diphones of all the possible combinations was used (those presenting higher frequency of occurrence).

2.3. Formant trajectories coding

By means of SRI's Decipher labels, the formant frequencies previously extracted are bounded in time for phone and diphone units (see Figure 2), yielding a matrix of 3 frequency values x #frames/unit for each linguistic unit. This variable-length segment is duration equalized to a number of frames equivalent to 250 ms, following results in

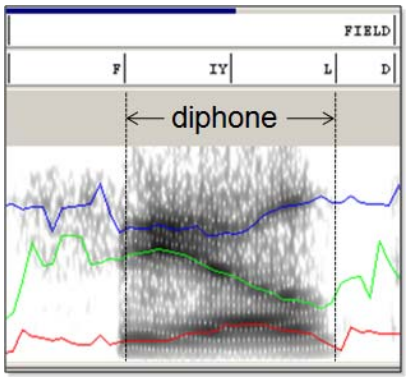

Figure 2: Detail of formant trajectories in a diphone unit. First three formant frequencies are shown.

previous studies [3][10]. Finally, those three temporal trajectories within a unit are coded by means of a fifth order DCT and its coefficients concatenated, yielding our final 3x5=15 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 Gaussian Mixture Model – Universal Background Model (GMM-UBM) framework [12], using duration-equalized DCT-coded formant frequency 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 Joint Factor Analysis (JFA) [8] and total variability modeling [2] techniques, which have outperformed the

former one 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 GMM-UBM 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.

UBM and *maximum-a-posteriori* (MAP) adapted speaker models [12] were trained and tested on unit-dependent data (using every unit segment available in both training and testing utterances), yielding an independent GMM-UBM system for each linguistic unit. In the case of phone units, 32 mixtures were used, and 8 mixtures in the case of diphone units. This procedure yields N scores per trial (N=#units) which can be used either as individual speaker recognition systems or, additionally, combined in a single fused system. None of these individual systems include any type of score normalization.

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 informative likelihood ratios for very short speech samples, as it is the case of forensic applications where a speech expert, typically a linguist/phonetician, can isolate or mark segments of compatible/comparable sounds between speech samples (typically, several segments belonging to some linguistic unit). On the other hand, when different types of information can be used, individual units are combined to achieve better discriminative capabilities.

Individual systems were combined in both intra- (different phones or different diphones between them) and inter-unit (pooling phones and diphones together) manners. Two different fusion techniques were used: sum fusion and logistic regression fusion. The former one was performed after linear logistic regression calibration per unit, 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 Equal Error Rate (EER) of the units to be fused, leaving out those performing worse. However, this procedure does not guarantee 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 an exhaustive 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 obtain calibrated LRs.

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. 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.

Phone unit	EER (%)	minDCF	C_{llr}	$minC_{llr}$
AY	21.67	0.0907	0.6949	0.6593
L	23.74	0.0966	0.7490	0.7173
AE	24.92	0.0922	0.7466	0.7161
R	25.47	0.0957	0.7672	0.7430
Y	26.03	0.0948	0.7916	0.7615
N	26.27	0.0942	0.7790	0.7554
AX	26.54	0.0990	0.8080	0.7750
PUH	27.36	0.0931	0.7925	0.7689
OW	27.78	0.0944	0.8088	0.7898
IH	28.92	0.0990	0.8172	0.7889

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

5.2. Phone units: individual and combined systems performances

Table 1 shows individual performance of the ten best performing phone units in terms of EER for the NIST SRE 2006 English-only male 1side-1side task. Although the performance of these phone-dependent systems is far from that of our reference system, it is actually a remarkable result taking into account the amount of speech used by each system (e.g., 5.9 seconds per utterance in average for the best performing phone 'AY'). Moreover, all of them have good calibration properties (small difference between Cllr and minCllr). This allows us to obtain informative calibrated likelihood ratios from very short speech samples (just the speech segments belonging to that unit present in the utterance), as we can see in the Tippett plot in Figure 3 for the best performing phone unit ('AY'). Furthermore, it has to be taken into account that the feature extraction process is affected by errors both in the formant tracking and in the ASR system (time alignment and phone decoding errors).

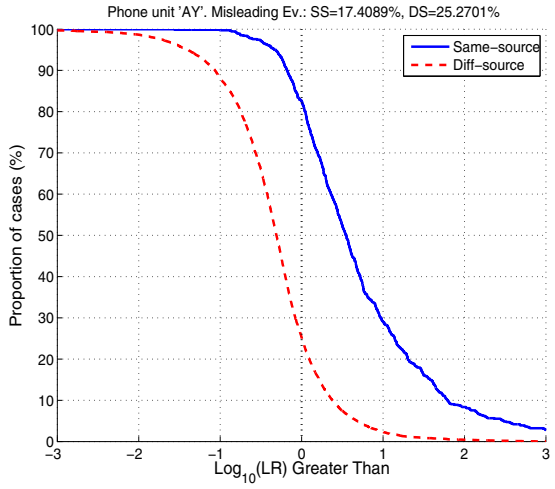

Figure 3: Tippett plot for the best performing phone unit ('AY') in the NIST SRE 2006 English-only male 1side-1side task.

Additionally, there are lots of units that can be combined to improve the discriminative power of the whole system. Figure 4 show the EER of the fused system as a function of the number of phones fused by means of the sum rule and logistic regression techniques, and for the two types of unit selection schemes for both techniques. Solid-line curves represent fusion results for different thresholds set for the EER, while circles represent the result for the unit selection algorithm. The performance of the reference system is also shown as a red dashed line. For the case of fusing phones performing better than a certain threshold, it can be seen that, for both type of fusion techniques, the EER of the fused system converge for a number of fused phones greater than 12, being this EER lower for the logistic regression technique (13%) than for the sum fusion rule (14%). In both cases, the performance of the fused system is greatly improved with respect to any of the individual phone systems, and quite close to that of the reference system (10.62%) using a much smaller amount of speech data (about 10% of the whole utterance for the case of fusing 12 phones). Moreover, it is worth noting that the unit selection algorithm used can achieve better fusion results (12.23%) than simply setting a threshold for the EER of the units to be fused in the case of the sum fusion rule.

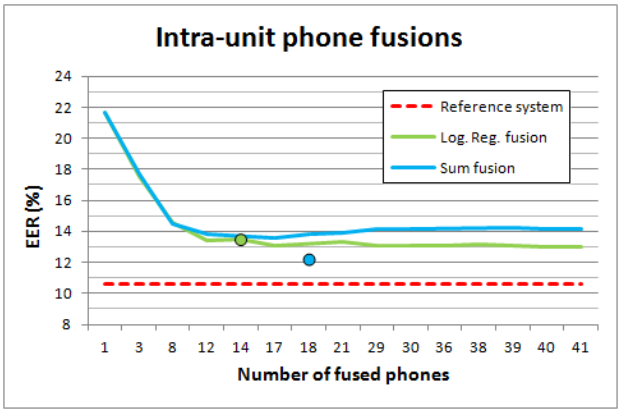

Figure 4: EER as a function of the number of phones combined (solid-line curves for combinations based on EER threshold, circles for the unit selection algorithm).

Diphone unit	EER (%)	minDCF	C_{llr}	$minC_{llr}$
Y-AE	29.65	0.0964	0.8269	0.8043
Y-UW	29.78	0.0993	0.8439	0.824
L-AY	30.46	0.0969	0.8343	0.8089
DH-AE	31.13	0.0980	0.8668	0.8413
AX-N	31.54	0.0992	0.876	0.8528
UW-N	31.67	0.0957	0.8634	0.8421
N-OW	32.92	0.0996	0.8738	0.8594
AE-N	34.86	0.1000	0.9024	0.8767
N-D	35.05	0.0995	0.9065	0.8884
L-IY	35.58	0.0995	0.9002	0.8822

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.

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 is a consequence of the feature extraction process and the generative modeling technique used; a particular two phone combination (diphone) has a fewer number of tokens in a speech sample than its constituent phones, and because we are coding each linguistic unit in a single feature vector, much less feature vectors are available to train the GMM of that unit. However, it can be seen that good calibration properties can be achieved, as in the case of phone units.

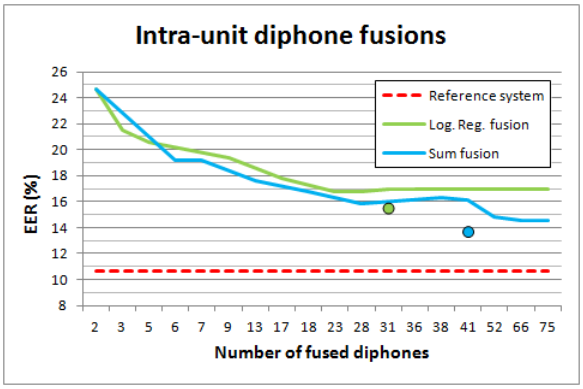

Figure 5: EER as a function of the number of diphones combined (solid-line curves for combinations based on EER threshold, circles for the unit selection algorithm).

Figure 5 show the results of the same experiments shown in Figure 4 but carried out with diphone units. In this case, the EER of the fused system converges for a higher number of fused units, and this EER is higher for logistic regression (17%) than for the sum rule (14.5%). Again, the unit selection algorithm achieves the better result for the sum fusion rule (13.7%).

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 6.

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 a 11.97% EER for the logistic regression fusion technique combining a high number of linguistic units (90). For the sum fusion rule, although the EER converges to a higher value, the unit selection algorithm can achieve again a better result (12.18%) with a reduced number of fused units (17).

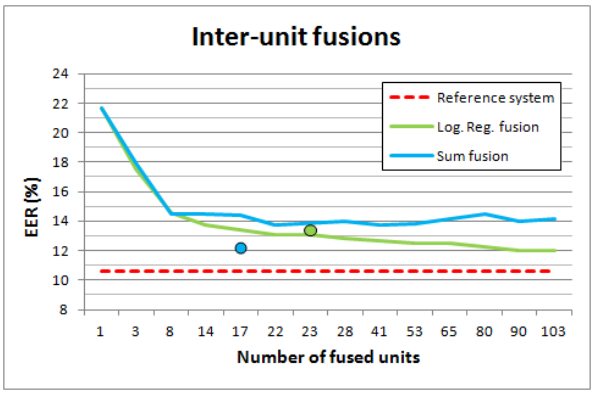

Figure 6: EER as a function of the number of units (phones+diphones) combined (solid-line curves for combinations based on EER threshold, circles for the unit selection algorithm).

5.5. Combination of formant trajectories and standard MFCC systems

For the best performing system based on formant trajectories in linguistic units (logistic regression fusion of 90 phones+diphones), sum fusions were carried out with both the reference GMM-UBM and an i-vector system [2] with Linear Discriminant Anaylisis (LDA), Within-Class Covariance Normalization (WCCN) and cosine distance scoring. This i-vector system is based on the same UBM as the reference GMM-UBM system, and total variability (400 dimensions), LDA (200 dimensions) and WCCN matrices were trained on one half of SRE05 dataset, while the other half was used for calibration purposes. In the case of the reference GMM-UBM system, the whole SRE05 dataset was used for calibration purposes, as it was done with the formant trajectories-based unit-dependent systems.

Table 3 shows performances for these three systems and the sum fusions between the two MFCC-based and the best fusion of formant-trajectories based systems. It can be seen that relative improvements of 26.8% and 17.2% on EER can be achieved for the GMM-UBM and the i-vector systems respectively, due to the different nature and time span of the formant-trajectories features.

System	EER (%)	minDCF
1) GMM-UBM MFCC	10.26	0.0457
2) i-vector MFCC	8.86	0.0407
3) Formant trajectories - best fusion	11.97	0.0636
Sum fusion of 1 and 3	7.51	0.0437
Sum fusion of 2 and 3	7.33	0.0356

Table 3. EER (%) and minDCF for MFCC-based systems, for the best system based on formant trajectories, and for the fusions between them, in the NIST SRE 2006 English-only male 1side-1side task.

6. Summary and conclusions

In this paper we have presented a higher level system based on formant trajectories that achieve remarkable performance in a very challenging speaker recognition task, exploiting the combination of multiple pieces of information distributed among the linguistic units under analysis.

While individual unit-system performances are far from that of MFCC-based systems, it has to be taken into account the much smaller amount of speech data used by them (less than 10 seconds per utterance for most of the units). Moreover, LRs provided by them are still very useful in applications such as forensic speaker verification, where the analysis is usually focused on linguistic features and formant trajectories can be used. In this way it is possible to deal with uncontrolled scenarios where only some short segments (compatible sounds between speech samples) are available to be compared, making it possible to inform about the speaker identity in the speech sample.

Furthermore 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 achieve a performance close to the results obtained with the same system framework based on MFCC features. Finally, it has been shown that the combination of this kind of systems with the state-of-the-art ones provides a significant improvement in the performance of the overall speaker recognition process.

References

[1] N. Brummer, et al. "Application-independent evaluation of speaker detection". Computer Speech and Language, (20) 230-275, 2006.

[2] N. Dehak, et al. "Front-End Factor Analysis for Speaker Verification". IEEE Transactions on Audio, Speech and Language Processing, 19(4), 788-798, May 2011.

[3] A. d. Castro, D. Ramos, and J. Gonzalez-Rodriguez. "Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking". Proc. of Interspeech'09, pp. 2343-2346, September 2009.

[4] J. Franco-Pedroso, F. Espinoza-Cuadros, and J. Gonzalez-Rodriguez. "Cepstral Trajectories in Linguistic Units for Text-Independent Speaker Recognition". IberSPEECH, volume 328 of Communications in Computer and Information Science, pp. 20-29, Springer, 2012.

[5] J. Franco-Pedroso, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, and D. Ramos. "Fine-grained automatic speaker recognition using cepstral trajectories in phone units". Quantitative approaches to problems in linguistics – Studies in honor of Phil Rose. Cathryn Donohue, Shunichi Ishihara, William Steed (editors). ISBN 9783862883844. LINCOM Studies in Phonetics 08, pp. 185-196, 2012.

[6] J. Gonzalez-Rodriguez 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.

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

[8] P. Kenny et al. "A Study of Inter-speaker Variability in Speaker Verification". IEEE Trans. on Audio, Speech and Lang. Proc., 16(5):980-988, 2008.

[9] K. McDougall. "Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies". International Journal of Speech Language and the Law 13(1), pp. 89-126, 2006.

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

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

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. "Speaker verification using adapted gaussian mixture models". Digital Signal Processing 10, pp. 19-41, 2000.

[13] P. Rose. "Forensic Speaker Identification". Taylor & Francis Ed., Forensic Science Series, 2002.

[14] E. Shriberg. "Higher-level features in speaker recognition". Speaker Classification I: Fundamentals, Features and Methods. C. Müller (Ed.), Lectures Notes in Artificial Intelligence 4343, pp. 241-259, Springer, 2007.

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

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

[17] Wikipedia contributors. "Arpabet". *Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/wiki/Arpabet

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