A New Phono-Articulatory Feature Representation for Language Identification in a Discriminative Framework

Nueva representación de características fono-articulatorias para identificación del idioma en un marco discriminativo

Oneisys Núñez Cuadra and José Ramón Calvo de Lara

Centro de Aplicaciones de Tecnologías de Avanzada,
Cuba
oneysita@yahoo.com, jcalvo@cenatav.co.cu

Article received on March 18, 2011; accepted on June 30, 2011

Abstract. State of the Art language identification methods are based on acoustic or phonetic features. Recently, phono-articulatory features have been included as a new speech characteristic that conveys language information. Authors propose a new phono-articulatory representation of speech in a discriminative framework to identify languages. This simple representation shows good results discriminating between English and Spanish, using a reduced training set of phono-articulatory trigrams vectors.

Keywords. Phonetic features, articulatory features, language recognition and support vector machines.

Resumen. Los sistemas de identificación de idiomas en el estado del arte se basan en características acústicas o fonéticas. Recientemente, las características fono-articulatorias han sido incluidas como una nueva caracterización del habla que contiene información sobre el idioma. Los autores proponen una nueva representación fono-articulatoria del habla usando un marco discriminativo para identificar idiomas. Esta simple representación muestra buenos resultados en la discriminación entre inglés y español, usando un reducido conjunto de entrenamiento basado en vectores de trigramas fono-articulatorios.

Palabras clave. Características fonéticas, rasgos articulatorios, el reconocimiento del lenguaje y las máquinas de vectores soporte.

1 Introduction

Automatic language identification (LID) systems are usually divided into two core areas: Front-end refers to the capture of features and back-end, refers to the classification of language models.

In the first area, the most used methods have been mainly supported by spectral (Torres-Carrasquillo, Singer et al. 2002; Singer, P.A. Torres-Carrasquillo et al. 2003) and phonotactic features (Singer, P.A. Torres-Carrasquillo et al. 2003; Glembek, Matějka et al. 2008). Purely spectral LID aims at capturing the essential differences between languages by modeling distributions of spectral features directly. This is typically done by extracting a set of spectral features from segments of speech. Moreover, systems based on phonotactic characteristic aim at incorporating information directly related to the language at different levels, depending on phonetic and/or word decoders, and in some cases more advanced lexical information.

In the second area, discriminative (Singer, P.A. Torres-Carrasquillo et al. 2003) or probabilistic (Torres-Carrasquillo, Singer et al. 2002) classification frameworks are used to identify the language-dependent patterns in such features.

These methods have been evaluated regularly during NIST[1] Language Recognition Evaluations in language detection tasks in conversational telephony under challenging channel conditions, the best system performance on 3, 10 and 30 second trials were 9 , 3 an 2 % respectively (NIST' 2009). The main drawback

[1] NIST: National Institute of Standards and Technology (USA), NIST is the federal technology agency that works with industry to develop and apply technology, measurements, and standards.

of all these methods is the huge amount of data necessary to train and test the classifiers and its intrinsically high computational cost, taking into account that these results were obtained by systems using a fusion of classifiers as backend.

Recently, some authors have proposed the use of sequences of phono-articulatory features (phono-AFs) for LID in a probabilistic framework (Parandekar and Kirchhoff 2003; Kanokphara and Carson-Berndsen 2006), modeling one or many sequences of sub-phonemic events such as articulatory classes like place, manner of articulation, voicing, rounding of lips, etc. A feature stream approach might also be of benefit for LID an improvements may be expected concerning the amount of training material, the length of the test material and its computational cost.

This feature-based approach has a range of potential advantages for LID (Parandekar and Kirchhoff 2003):

1. The number of articulatory classes in any given language is typically much smaller than the number of phone classes; training data features can be shared across phones, leading to a larger number of training samples per class. Models can therefore be trained more robustly than the corresponding phone models. Moreover, the overall number of models is reduced.
2. The language-independent nature of phono-articulatory features enhances the portability of the system to new languages or dialects.
3. A large part of cross-linguistic variation arises from differences in articulatory timing; such phenomena may have language-differentiating potential. (Stüker, Metze et al. 2003).
4. Articulatory movements are much less affected by the environmental conditions (Kirchhoff 1999) than their acoustic representations.

Our proposal to use a new representation of phono-AFs in LID is based on the fact that:

- The dynamic relation between each phono-AFs stream and its probability of co-occurrence brings information about the language (Parandekar and Kirchhoff 2003)
- The use of sequences of phono-based trigrams are well established in LID phonotactic approaches (NIST'2009 ; Singer, P.A. Torres-Carrasquillo et al. 2003)
- Discriminative classification methods as SVM are used in spectral LID (NIST'2009 ; Singer, P.A. Torres-Carrasquillo et al. 2003)

Then if, for a sequence of utterances of a language, streams of phono-AFs are combined to obtain AFs vectors, codified adequately and concatenated in a phono-based trigram, a new representation of phono-AFs will be obtained which reflects the short-time dynamic of the articulators. These trigrams of phono-AFs vectors can be classified in a discriminative framework in order to identify a target language from an impostor language with a reduced computational cost and suitable efficacy.

From now on, this paper is organized as follows. Section 2 explains the representation of phono-AFs of speech. Section 3 describes our proposal. The LID experiment was described in Section 4. Section 5 discusses the results and Section 6 offers conclusions and suggests future work.

2 Phono Articulatory Representation of Speech

If we take into consideration that utterances can be modeled as a sequence of phonemes, and correspondent phones are generated by a determined configuration of vocal tract articulators, each phoneme can be mapped to its own articulatory configuration.

The term “articulatory” does not refer to the real estimation of vocal tract parameters from the acoustic waveform. While the integration of real physiological measurements with spoken systems is desirable, it is at the same time impractical since this kind of data is not abundant (Only EMA Database). (Wrench 1999)

One specific means to achieve a relatively accurate phonetic characterization of the speech signal is through the use of AFs instead of phonetic segments. These abstract classes describe important vocal tract properties and articulatory motions during speech production and characterize the most essential aspects of articulatory properties of speech sounds (e.g., phonation, nasality, roundedness, etc.) in a quantized form, leading to an intermediate representation between the speech signal and the lexical units (Kirchhoff 1999).

This AF representation could be considered as a simplified alternative to phonological articulation for many purposes and, as a compact speech representation, it could replace the phone sequence of an utterance for a parallel AF representation, assuming that this combined information completely specifies the acoustic speech signal.

There are many advantages of using AFs. AFs based approaches can be more accurate when discriminating phones that can ‘‘sound similar’’ but differ significantly in production. Furthermore, it may be able to explicitly capture changes in articulation due to the differences in language. These can be useful cues that AFs based systems can exploit. In addition, it proved to be more robust in noisy environment. (Kirchhoff, Fink et al. 2002)

In a phonetic transcription of speech, (the one proposed by IPA (IPA 1999) is the most widely used) differences between phonemes are annotated as different articulatory configurations. Such transcriptions can be considered as annotating details of the articulatory structure of each phone and reflect the temporal organization of vocal tract articulators. That is why this representation is named “phono-articulatory”.

Phono–AFs are grouped to be maximally distinct to the acoustic domain of phone realization. Then, the most effective classification method for achieving high phone representativeness will group those features as values of a single articulatory class which display maximal acoustic contrast, for example: manner and place of articulation, phonation, rounding and position of the tongue (Kirchhoff, Fink et al. 2002; Kirchhoff 1999).

- Manner of articulation refers to the way the articulation is accomplished.
- Place of articulation is the place where the articulators obstruct the air.
- Speech sound phonation depends on the state of the vocal cords and constriction in the oral tract.
- Rounding feature describes whether the sounds are generated when the lips are rounded or not
- Speech sound characteristics also strongly depend on the shape and the position of the tongue, and its place of articulation: front/back.

For example, consonants can be described by their manner and place of articulation, and their phonation. On the other hand, vowels can be described by their roundedness, height and backness. Then, phono-AFs can be represented as a feature vector where each feature value corresponds to an articulatory class.

3 Language Identification using Articulatory Features: Our Proposal

Streams of AFs have been used for LID, in a probabilistic framework, (Parandekar and Kirchhoff 2003; Kanokphara and Carson-Berndsen 2006) modeling sequences of sub-phonemic events such as articulatory classes e.g. manner , place of articulation, phonation, roundedness, etc. Methods for fast LID using short test signals have been proposed, considering that a multiple phonetic feature stream approach might also be of benefit for LID, expecting improvements with respect to the amount of training material, the length of the acoustic test signal, memory requirements and accuracy. They showed that such a multi-stream feature-based system significantly outperforms a comparable phone-based system while using fewer parameters.

Taking into account these previous results, our proposal is to reduce even more the computational cost of LID using phono-AF vectors instead of streams, but in a discriminative environment.

The proposed articulatory classes and corresponding feature values are shown in Table 1. For a sequence of phonetic labeled training utterances of two languages (English and Spanish), sequences of phono-AFs vectors using IPA look-up table are obtained, these vectors are concatenated in a trigram based configuration, and these trigrams of phono-AF vectors are used to train a SVM classifier.

Afterwards, by using a similar procedure for testing utterances in both languages, the classifier is evaluated in order to identify a target language (Spanish) from an impostor language (English) with a reduced computational cost and a suitable efficacy.

Table 1. Articulatory classes and possible values of features

Articulatory Classes	Possible values of features
Manner	Stop, affricate, fricative, nasal, lateral, approximant, flap, trill, vowel tense, vowel lax, sil
Place	bilabial, labio-dental, dental, alveolar, post-alveolar, palatal, velar, glottal, retroflex, (*front, central, back*)[2], sil
Phonation	voiced (+) , voiceless (-), sil
Height	Open, mid, close, nil , sil
Rounding	+round, -round ,nil, sil

4 Language Identification Experiment

Sequences of phonetic labels of English and Spanish utterances were obtained from OGI Multilanguage telephonic speech Database V1.2 (Muthusamy, Cole et al. 1992). Phonetic transcriptions of speech were hand generated and manually time aligned, so we can assume that respective phono-AFs are fine time aligned, representing authentically each phonetic label.

During the phono-articulatory transcription process, the phonetic transcriptions were mapped automatically to the code of its correspondent phono-AF vector by a homemade system using a look up table that links the phonemes with its articulatory characteristics defined by IPA. The look-up phoneme to articulatory table, using IPA chart for both languages, is summarized in Table 2.

Phonetic realizations not labeled by IPA but present in the Database, were labeled using Worldbet Multilanguage extended IPA alphabet. Each one of the possible values for each articulatory class were adequately codified, assigning 0 for values “nil” and “sil” for all classes, and values {1,2} for Phonation and Roundedness, {1,..,10} for Manner, {1,..,12} for Place, {1,.,3} for Height and {1,..,4} for Extras. The “Extra” class reflects other properties of the hand labeled phonemes, not included in any articulatory class, for example the influence of the diacritics.

So, each transcribed phoneme is codified by a 6 dimensional vector and the phonetic-based trigrams sequences of these vectors (18–dimesional vectors) were used for the LID experiment in SVM framework. Figure 1 shows our experimental scheme.

Many distinctive phono-AF vectors were obtained for each language (see Table 2), 21 for English, 17 for Spanish and only 17 in common, giving a previous idea that the LID experiment results with a discriminative classifier could get satisfactory outcomes despite having a considerable percentage of phono-AFs in common (50% for Spanish).

Moreover, the main idea is to use trigram configuration taking into account that these new vectors represent, in some way, the phone stream dependencies of the languages. Many different trigrams combinations like “tokens” were obtained. Also several experiments were performed to detect which of them were valid (usually corresponding to syllable nuclei) and to what extent these combinations are repeated in each language in question. This phase of experimentation allowed us to detect the enormous amount of redundant information contained in the training data and how these languages were separable given the proposed representation.

[2] Front, central and back are the values corresponding to the Front/Back articulatory class for vowels.

Table 2. Look-up phoneme to articulatory table for English and Spanish

	IPA	Worldbet	Articulatory classes					
			Manner	Place	Voicing	Rounding	Vertical	Extras
Phono-articulatory features common to both languages	b	b	Stop	Bilabial	+	nil	nil	nil
	k	k	Stop	Velar	-	nil	nil	nil
	g	g	Stop	Velar	+	nil	nil	nil
	ʧ	tS	Affricate	Postalveolar	-	nil	nil	nil
	ɟʝ	dZ	Affricate	Postalveolar lar	+	nil	nil	nil
	f	f	Fricative	Labiodental	-	nil	nil	nil
	θ	T	Fricative	Dental	-	nil	nil	nil
	ð	D	Fricative	Dental	+	nil	nil	nil
	s	s	Fricative	Alveolar	-	nil	nil	nil
	m	m	Nasal	Bilabial	+	nil	nil	nil
	ŋ	N	Nasal	Velar	+	nil	nil	nil
	w	w	Aproximant	Velar	+	nil	nil	nil
	j	j	Aproximant	Palatal	+	nil	nil	nil
	i	i	Vowel tense	Front	+	-	Close	Long
	ɪ	I	Vowel lax	Front	+	-	Close	Short
	u	u	Vowel tense	Back	+	+	Close	nil
	ə	&	Vowel lax	Central	+	+	Mid	Short
	ɛ	E	Vowel lax	Front	+	-	Mid	Short
	ɔ	o	Vowel tense	Back	+	+	Mid	nil
	ʌ	^	Vowel tense	Central	+	-	Mid	nil
English Phono-articulatory features	p	ph	Stop	Bilabial	-	nil	nil	Aspirated
	t	th	Stop	Alveolar	-	nil	nil	Aspirated
	d	d	Stop	Alveolar	+	nil	nil	nil
	k	kh	Stop	Velar	-	nil	nil	Aspirated
	v	v	Fricative	Labiodental	+	nil	nil	nil
	s	s	Fricative	Alveolar	-	nil	nil	nil
	ʃ	S	Fricative	Postalveolar	-	nil	nil	nil
	ʒ	Z	Fricative	Postalveolar	+	nil	nil	nil
	h	h	Fricative	Glotal	-	nil	nil	nil
		m=	Nasal	Bilabial	+	nil	nil	Syllabic
	n	n	Nasal	Alveolar	+	nil	nil	nil
		n=	Nasal	Alveolar	+	nil	nil	Syllabic
		N=	Nasal	Velar	+	nil	nil	Syllabic
	l	l	Lateral	Alveolar	+	nil	nil	nil
	ɹ	9r	Aproximant	Retroflex	+	nil	nil	nil
	ʊ	U	Vowel lax	Back	+	+	Close	Short
	ɜ	3r	Vowel lax	Central	+	-	Mid	nil
		&r	Vowel lax	Central	+	-	Mid	Short
	æ	@	Vowel tense	Front	+	-	Open	Long
	a	A	Vowel tense	Back	+	-	Open	nil
	ɒ	5	Vowel lax	Back	+	+	Open	nil

	IPA	Worldbet	Articulatory classes					
			Manner	Place	Voicing	Rounding	Vertical	Extras
Spanish Phono-articulatory features	p	p	Stop	Bilabial	-	nil	nil	nil
	t	t[	Stop	Dental	-	nil	nil	nil
	ð	D	Fricative	Dental	+	nil	nil	nil
	x	x	Fricative	Velar	-	nil	nil	nil
	χ	G	Fricative	Velar	+	nil	nil	nil
	β	V	Fricative	Bilabial	+	nil	nil	nil
	z	z	Fricative	Dental	+	nil	nil	nil
	ʝ	Z	Fricative	Palatal	-	nil	nil	nil
	ɲ	n~	Nasal	Palatal	+	nil	nil	nil
	n	n	Nasal	Dental	+	nil	nil	nil
	l	l	Lateral	Alveolar	+	nil	nil	nil
	ʎ	L	Aproximant	Palatal	+	nil	nil	nil
	ɾ	r(	Flap	Alveolar	-	nil	nil	nil
	r	r	Trill	Alveolar	+	nil	nil	nil
		lx	Vowel lax	Back	+	-	Mid	Long
	e	e	Vowel tense	Front	+	-	Mid	Short
		a	Vowel tense	Central	+	-	Open	nil

4.1 Reduction of Redundant Trigrams to Train the SVM

Phonetic transcriptions of 83 Spanish expression utterances as target, and 87 English expression utterances as non-target, were taken as training data from OGI Database, with just duration of 51 and 59 minutes respectively. NIST evaluations provide approximately 1 ½ hours of signals per language in addition to the training data of the systems (NIST'2009).

As we saw above, a considerable amount of redundant information was obtained as result of the phono-AF transcription and trigrams configuration (see Table 3 column "Before"). So, an important step on the experiment is to reduce this redundant information in the training data since this does not provide new information, but rather affects the efficiency given the discriminative nature of the SVM classifier.

Considering that, in order to train a SVM classifier only a set of representative feature vectors for each training class is required, a previous reduction of trigrams of phono-AFs vectors was done. It consists in selecting only one sample of each different trigram (vector)

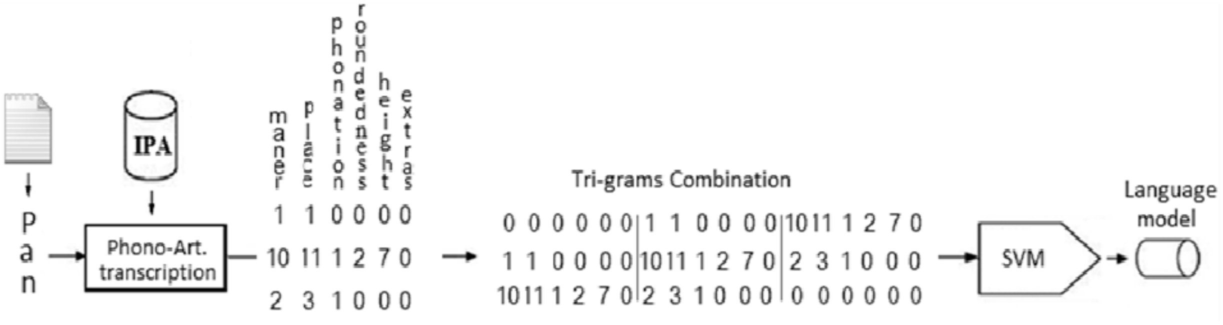

Fig.1. Scheme of the Phono-AF LID proposed system

for each language, causing a marked reduction of training data.

Table 3. Number of total trigrams before and after the trigrams reduction

Language	Before	After	%
English	461341	13408	2.9
Spanish	411642	7393	1.8

The total numbers of trigrams and the number of selected different trigrams are shown in Table 3. Observe the huge reduction of training data (less than to 3 % of total).

As a result of this analysis a new training set was obtained consisting of a 63.2% own combinations of English, 34.3% Spanish-specific and 1.24% in common. This unbalance between the selected training data for each language, could provoke a biased classification result.

4.2 SVM classifier for Language identification experiments

SVM is a binary maximum margin classifier, which by using a kernel function of the sequence of data to be classified can produce complex decision functions without a large amount of training data. Given two separable classes, the decision boundary is found by maximizing the margin between the support vectors (SV) such that no data occupies the space in-between. When the data is non-separable a soft margin is used. Incursions of data into the margin are penalized so a search for the best solution maximizes the margin and minimizes the penalties simultaneously. The trade-off between both is controlled by a single regularization parameter, C.

An exploratory analysis of training data was done in order to evaluate its distribution in the space of the proposed representation. A Principal Component Analysis (see Figure 2) showed that there is no linear separation between the representations of the two languages.

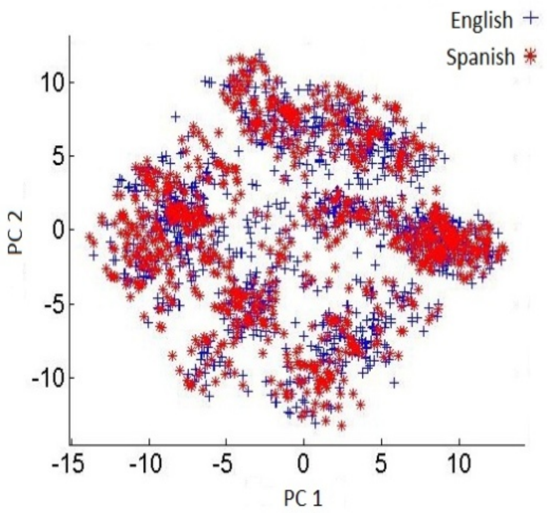

Fig.2. Principal Components Analysis was done with 500 trigrams representatives of the both languages. The red plus marks correspond to Spanish and the blue asterisk marks to English

So, SVMTorch (Collobert and Bengio 2001) with a RBF kernel was used for LID experiments given the nonlinear separability of the data. A cross-validation process with 10 folds was used for selecting the hyperparameters.

5 Experiment Results

A set of 1063 randomly selected short expressions, not contained within the training set, were taken for the testing process. All test utterances were less than 10 sec in duration and only 120 of them were over 3 sec.

A total of 521 test expressions were from Spanish (Pos.) and 542 for English (Neg.).

The results of LID experiments using trigrams of phono-FAs are shown in Table 4. It is remarkable that all classification errors identified were driven by expressions with very short duration (1.5 sec on average).

Table 4. Results of language identification experiments

Test Utt.	Error (%)	Pos.	F Neg.	Neg.	F Pos.
1063	3.1 %	488	33	542	0

As we assumed in Section 4, the experiment with manual transcription reflects excellent IER[3] results, because the phonetic transcriptions of speech were hand generated and manually time aligned.

Besides, English IER results (Neg.) were excellent but biased (not FPos.), since selected training trigrams for English are much more than for Spanish (See Table 3 column "After").

6 Conclusions and Future Work

The LID experiment results are very good. This new phono-AFs representation of speech has the following advantages over other features commonly used in LID methods:

- A reduced set of training utterances (les than one hour for each language) with an additional considerable reduction of trigrams features, with a correspondent reduction in computational cost.
- A mean duration of test utterances is about or less than 3 sec., a high challenging condition for LID.

This paper reflects that the use of vectors of phono-AFs concatenated in a trigram-based, could be considered as a new language discriminative feature with a reduced computational cost in front of short duration of test utterances.

However, we consider that the results are subject to the source of phono-AFs, not "fully automatic" obtained and fitted to each language. Future work will be, in order to eliminate these limitations of the experiments, to obtain AFs from automatic "acoustic to articulatory" mapping, getting rid of the influence of phonetic transcriptions of speech and achieve language independence and a fully automatic method.

Because of differences in acoustic domain of phone realization for different accent, possible influence of language accent must be evaluated too.

[3] IER: Identification Error Rate, one performance measurement of biometric identification methods.

References

1. **Burges, C. J. C. (1998).** A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery, 2(2), 121-167*
2. **Collobert, R. & Bengio, S. (2001).** SVMTorch: Support vector machines for large-scale regression problems. *The Journal of Machine Learning Research,* 1(9), 143-160.
3. **Glembek, O., Matejka, P., Burget, L. & Mikolov, T. (2008).** Advances in Phonotactic Language Recognition. *Interspeech 2008,* Brisbane, Australia, 743-746
4. **International Phonetic Association (1999).** *Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet.* Cambridge, U.K.: Cambridge University Press.
5. **Kanokphara, S. & Carson-Berndsen, J. (2006).** Articulatory-Acoustic-Feature-based Automatic Language Identification. *ISCA Workshop on Multiingual Speech and Language Processing (MULTILING 2006)*, Stellenbosch, South Africa.
6. **Kirchhoff, K. (1999).** *Robust Speech Recognition Using Articulatory Information.* Ph.D. Thesis, Universitat Bielefeld, Bielefeld, Germany.
7. **Kirchhoff, K., Fink, G. A. & Sagerer, G (2002).** Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication,* 37 (3-4) 303-319.
8. **Muthusamy, Y. K., Cole, R. A. & Oshika, B. T. (1992).** The OGI multilanguage telephone speech corpus. *International Conference on Spoken Language Processing*, Alberta, Canada, 895-898.
9. **Parandekar, S. & Kirchhoff, K. (2003).** Multistream language identification using data-driven dependency selection. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China, I-28- I-31
10. **Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M. & Reynolds, D.A. (2003).** Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification.*8th European Conference on Speech Communication and Technology (EUROSPEECH 2003),* Geneva, Switzerland, 1345-1348
11. **Stüker, S., Metze, F., Schultz, T. & Waibel, A. (2003).** Integrating Multilingual Articulatory Features into Speech Recognition. *8th European Conference on Speech Communication and Technolo-*

gy (EUROSPEECH 2003). Geneva, Switzerland, 1033-1036

12. The 2009 NIST Language Recognition Evaluation (August 11, 2009) Retrieved from http://www.itl.nist.gov/iad/mig//tests/lre/2009/lre09_eval_results/index.html.

13. **Torres-Carrasquillo, P. A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A. & Deller Jr., J. R. (2002).** Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features. 7th *International Conference on Spoken Language Processing,* Denver, CO, ISCA, 89-92.

14. **Wrench, A. (1999).** MOCHA-TIMIT Retrieved from http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html.

Oneisys Núñez Cuadra

*Graduada de Ingeniería Informática en el Instituto Superior Politécnico José Antonio Echeverría en julio del 2007.Trabajó en el Dpto. de Ingeniería de Sistemas del CENATAV desde esa fecha hasta el 2010 en que pasó a trabajar en una empresa. Alcanzó la Categoría Científica de Aspirante a Investigador. Sus intereses de investigación estaban en el Reconocimiento de Patrones, la Ingeniería de Software, el Análisis de Sistemas y el Procesamiento d*e Señales.

Jose Ramón Calvo de Lara

Graduado de Ingeniero en Telecomunicaciones en el Instituto Superior Politécnico José A. Echeverría en 1978.Obtuvo el título de Dr. en Ciencias Técnicas en el 2003 y la Categoría Científica de Investigador Titular en el 2004.Es investigador del Departamento de Reconocimiento de Patrones del CENATAV, donde dirige el grupo de procesamiento de voz y habla. Sus intereses de investigación están en la extracción, selección y modelación de rasgos de la voz para el reconocimiento del locutor, del lenguaje y del habla, los Métodos de enfrentamiento al ruido en la voz y la Detección robusta de actividad de voz.