UNIVERSIDAD DE VALLADOLID

ESCUELA TÉCNICA SUPERIOR

INGENIEROS DE TELECOMUNICACIÓN

TRABAJO FIN DE MASTER

MASTER UNIVERSITARIO EN INVESTIGACIÓN

EN TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES

Analysis of the GOP metric for assessing non-native Spanish pronunciation in the SAMPLE corpus

Autor:

D. Vandria Eunise Alvarez Alvarez

Tutores:

Dr. D. David Escudero Mancebo
Dr. D. Valentín Cardeñoso Payo

Valladolid, 9 de Julio de 2015

TITLE:	**Analysis of the GOP metric for assessing non-native Spanish pronunciation in the SAMPLE corpus**
AUTHOR:	**D. Vandria Eunise Alvarez Alvarez**
TUTORS:	**Dr. D. David Escudero Mancebo** **Dr. D. Valentín Cardeñoso Payo**
DEPARTAMENT:	Computer Science

Tribunal

PRESIDENTE:	**Dr. D. Pablo de la Fuente Redondo**
VOCAL:	**Dr. D. Carlos Vivaracho Pascual**
SECRETARIO:	**Dr. D. Juan Blas Prieto**
FECHA:	**9 de Julio de 2015**
CALIFICACIÓN:	

Abstract

This work presents an analysis over the set of results derived from the FGOP algorithm for the evaluation of pronunciation at phoneme level over the SAMPLE corpus. This corpus includes several recordings of uttered sentences by distinct speakers. These utterances have been transcripted with the help of a group of persons from the linguistic field collaborating with our research group. The results have been processed and analyzed to try identifiying possible improvements. Several observations over the metrics behavior are exposed. The phoneme dependence is discussed to suggest the establishment of thresholds that could enhance the metrics performance. Additionally, new scoring proposals are presented which are based on computing the loglikelihood values obtained from the FGOP algorithm and the application of a set of rules to obtain a new parameter that will allow to get a new score for every phoneme. With these new scores, a global score is generated to assess the pronunciation quality at speaker level. Finally, the global scores of every speaker are contrasted with the FGOP and human judgments scores.

Keywords

Computer Assisted Pronunciation Training (CAPT), Goodness of Pronunciation (GOP), Phonemes, Hidden Markov Models (HMM), Automatic Speech Recognition (ASR), Pronunciation

Resumen

Este trabajo consiste en el análisis de los resultados obtenidos en la evaluación de pronunciación a nivel fonema utilizando el algoritmo Forced GOP que ha sido imple-

mentado para ello. Se ha hecho uso de locuciones de diferentes oraciones realizadas por distintos hablantes, las cuales han sido grabadas y anotadas dentro del corpus SAMPLE. Este corpus fue desarrollado dentro de nuestro grupo de investigación en colaboración con personas del ámbito lingüista. Se ha trabajado con los datos obtenidos para identificar posibles mejoras, se hacen varias observaciones en el comportamiento de la métrica y se discute la dependencia a nivel fonema y hablante que sugiere el establecimiento de posibles umbrales para mejorar su rendimiento. Además se agregan propuestas en base a los datos de loglikelihood que arroja la FGOP y se aplican una serie de reglas para establecer un nuevo parámetro que permita dar una calificación por cada fonema. Estas calificaciones permiten generar una calificación global de pronunciación a nivel hablante. Las puntuaciones globales se han contrastado con los resultados de la FGOP y las evaluaciones realizadas por jueces humanos.

Palabras clave

Pronunciación Asistida por Computadora (CAPT), Goodness of Pronunciation(GOP), Fonemas, Modelos Ocultos de Markov, Reconocimiento Automático del Habla (ASR), Pronunciación.

Acknowledgements

I would like to start thanking the University of Valladolid and Banco Santander for their support through the scholarship program that granted me the opportunity to come to Spain and study this master. To my advisors David Escudero and Valentín Cardeñoso for their time and recommendations, without their assistance this work would not have been possible.

My special thanks to my parents Cesar and Miriam, for always believing in me, even when I did not. To Juan Luis, for being by my side and for the countless conversations giving me relief and support. To my family and friends in Guatemala for always demonstrating their unconditional support. To those special persons I met and became my friends and second family, thanks for making me feel closer to home. Finally, to my grandparents for always guiding me with the help of God.

"Limits exist only in the mind". *Aristotle.*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The increasing globalization has led to a higher demand for knowledge acquisition regardless of the place or language. There are several technological tools that help with daily activities when a foreign language is involved. Nowadays, is possible to understand a document posted on Internet (in another language) by using an on-line translator, watching a movie in a foreign language with subtitles, etc. However, people is aware of the necessity of learning a foreign language to be competitive in the market, or like in some cases, just because is a personal achievement.

Learning and teaching a foreign language is not an easy task for both the student and the teacher. From the student's point of view, sometimes the larger difficulties include the time dedication and location, since learning a language requires of periodical practices and assistance of an expert in a given place. Likewise, from the teacher's point of view there are also other important characteristics, such as, establishing an appropriate methodology based on the student's level, time disposition, whether it is possible or not to have personal classes and more importantly the student's native language [21].

The teaching of a foreign language concerns with aspects such as grammar, reading, listening and speech. Each of these is approached in distinct manner. In most of the language classes there is often a deficiency over the speech area, since this requires not only practice but the judgment of an expert telling the student his mistakes and how to improve them in a direct session.

There are several features that can be used to grade a student's speech, which can be related with the prosody or with the phonetics. Prosodic features look for a a general grading of the pronunciation, such as fluency, intelligibility, among others. On the other hand, the phonetic features evaluate on a lower level how well pronounced a phone was. The last, is a harder job because the phone discrimination requires a great knowledge of the language phonetics and very good listening capabilities.

1.1 *Motivation and problem description*

A few years ago I was doing an exam in a computer laboratory to get an English certificate that I needed for the university. The exam was divided into four categories, first some questions and vocabulary were asked and I just had to click the right answer. The same procedure for answering was used for the listening and reading categories. However, the speech category consisted of giving instantaneous spoken answers that were recorded to be evaluated afterwards by human judgments. Hence, to get my grade and certificate I had to wait for a month.

The listening and reading categories were automated so the grading was instantaneous, but the speech evaluation was not, therefore, the delay on delivering the final grade. If this category would have been also automated, then probably I would have get that same day my certificate. The latter is just one of the many examples on which there is a great necessity for automated systems that are capable to assess a person speech, whether if it is for an evaluation or just for practicing.

There are multiple features that can be used to measure the quality of pronunciation. These can be classified into phonemic and prosodic [25]. To realize an automated system capable of assessing a speaker's pronunciation is not an easy job. In fact, since there is no established framework of features that can be used during the evaluation, different paths have been taken to try giving a solution.

Utilization of Computer Assisted Pronunciation Training (CAPT) systems for pronunciation assessment started in the earliest 2000's and it has been rising parallel to the increasing advancements in technologies [22]. The majority of CAPT systems make use of Automatic Speech Recognition technologies. Usually the ASR-based CAPT systems incorporate distinct phases that evaluate different features of pronunciation.

The scoring phase evaluates the pronunciation quality by giving a score obtained from the comparison of temporal properties with the references (usually native speech). The error detection phase also deals with giving a score but on a lower level, like quality of phonemes pronunciation. To compute this scores there are several proposals over the state of art works.

Assessment at phoneme level often uses confidence measures, since these have easier implementations compared to others when using an ASR engine based on Hidden Markov Models [17]. A confidence measure quantifies how well a model fits the corresponding data and its efficiency is determined by correlating its results with the human judgments scores. There are various types of confidence measures, one of them is the Goodness of Pronunciation (GOP) metric first developed by Witt [25]. The GOP metric employs orthographic transcriptions to describe the phonemes sequence uttered and Hidden Markov Models to calculate the likelihood of a given acoustic segment with all possible phonemes.

Several research has been made based on the GOP metric to accomplish different purposes. These include assessment in CAPT systems with other languages (not English),

adaptations of the HMMs, establishment of different thresholds or changing the computation of the log-likelihoods, among others. The forced GOP is an example of these changes over the GOP original metric.

The forced GOP is a variant from the original, and it has been implemented by our research group to evaluate the SAMPLE corpus. This corpus comprehends a series of uttered sentences that are evaluated at phoneme level. The outcome of this evaluation is analyzed through this work to detect strengths and weaknesses in the GOP implemented.

1.2 *Objectives*

Previously the different paths, features and characteristics related to accomplishing an automated pronunciation assessment were described. Based on these, the objective of this work is established:

- **Analyze the GOP results over the SAMPLE corpus to detect possible weaknesses and therefore try to give new proposals that will improve the assessing of non-native Spanish pronunciation.**

To aboard the latter, distinct partial objectives have been studied. All these, included the context are represented in a blocks diagram in figure 1.1. The partial objectives are:

- *Conduct a review over the state of the art works that have described, designed and implemented the GOP metric.*

  To reach this objective we will start studying the first publications that explained the GOP algorithm, followed by looking the fundamentals on which this metric is based, to finally depict the most important aspects of some of the GOP implementation variants.

- *Contrast the GOP results over the SAMPLE corpus to identify its strengths and weaknesses.*

  To accomplish this, all the GOP results corpus are statistically analyzed at different levels (phoneme, speaker, groups of phonemes, etc.).

- *Study possible alternatives that also allow to give a phoneme and speaker scoring for pronunciation assessment.*

  To do this, different new metrics based on the GOP results corpus are computed, which will basically consist of applying a series of rules that rely on the logarithmic scores and the GOP score at phoneme level.

The realization of the previous objectives pretends to denote the strengths and weaknesses of the GOP implemented to score the SAMPLE corpus and therefore obtain possible new metrics that correlate well with the human judgments.

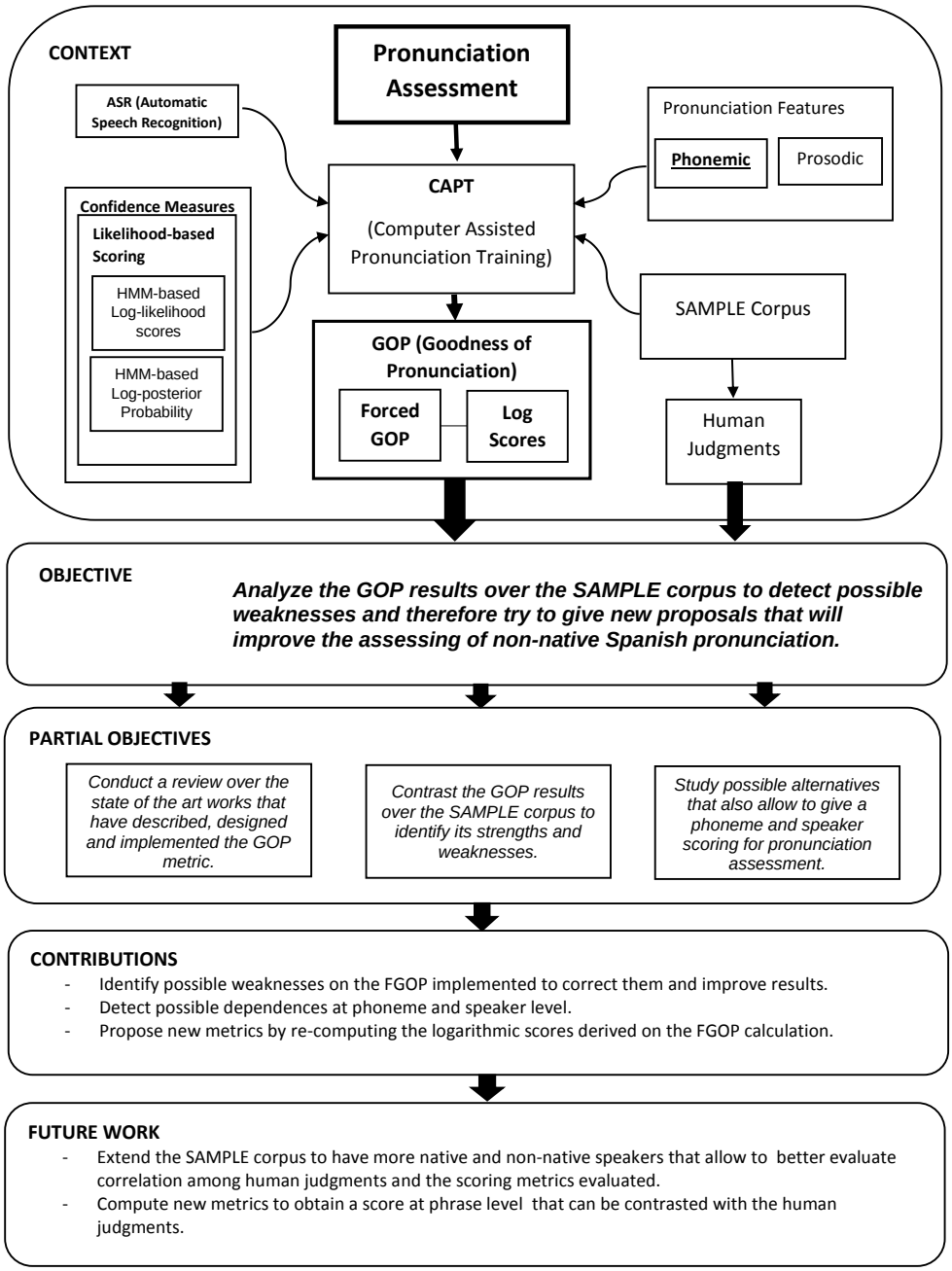


Figure 1.1: General scheme of the TFM, representation of context, objectives, contributions and future work.

1.3 *Methodology*

The methodology used in this work can be detailed as follows:

- A detailed study over the features and important characteristics related to the pronunciation assessment. Followed by a survey over the state of the art works that have used the GOP metric to assess the pronunciation quality.
- Computation and analysis of the data derived from evaluating a subset of the SAMPLE corpus with the GOP algorithm previously implemented.
- Creation of new scoring proposals that improve the correlation with the human judgments.

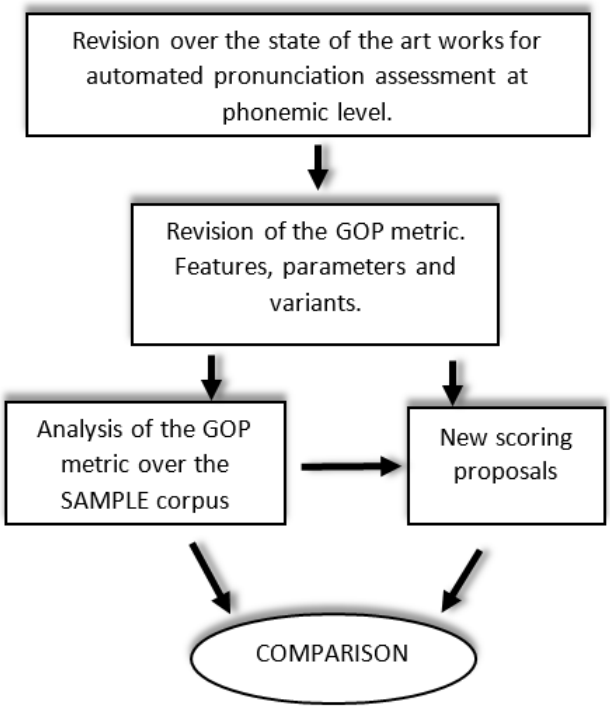


Figure 1.2: Methodology diagram

1.4 *Document structure*

This work is structured into 3 more chapters and an appendix:

- The Chapter 2 comprehends the different aspects related to the automated pronunciation assessment. First the language, phonetics and phonology most important features are depicted from a linguistic point of view. Second, a pronunciation assessment section is presented, which starts by describing some of the classifications

on pronunciation errors, followed by the description of the Computer Assisted Pronunciation Training (CAPT) and the use of Automatic Speech Recognition (ASR) technologies on them. Third, the error detection topic is extended and this time it is focused on the automated techniques used along the past years, such as confidence measures, the likelihood ratios and GOP. Finally, the GOP metric and the algorithm implementation is detailed and also some of the variants found in the literature.

- Chapter 3 presents the materials, results and discussions. First the different data utilized are detailed, to give the reader a perspective of the data types analyzed. Second, the statistical analysis over the GOP scores is showed and the different statistics are described and discussed. Third, each of the proposals for computing new pronunciation scores are showed to finally compare these with the human judgments.

- Chapter 4, presents the conclusions obtained in this work, its limitations, some recommendations and possible future work.

- Appendix A, contains the set of all the graphics from the various statistics computed over all speakers.

Chapter 2

State of the Art

2.1 Language, Phonetics and Phonology

2.1.1 *Langue and Parole*

There are two important aspects to distinguish in the language field proposed by the linguist Ferdinand de Saussure, the *langue* (language) and *parole* (speech). The langue refers to the general model that remains constant on all members from the same linguistic community, is the system of a language. Parole on the other side, is the speech itself, a realization or act of the langue in a given moment and place by a community member [3].

2.1.2 *Linguistic sign*

The *linguistic sign* is composed by the *significant* (expression) and the *meaning* (content,concept or idea). These two facets have different functionality over the langue and parole planes [16].

Components	Langue plane	Parole plane
Significant	Rules system that order the phonic aspects of the langue.	Group of sounds that can be perceived by human ear.
Meaning	Represented by abstract rules (syntactical, morphological, etc.)	Concrete communication, which only has sense in its totality.

Table 2.1: Linguistic sign components over the langue and parole planes

2.1.3 *Phonetics and Phonology*

Based on the definitions mentioned before the study of language sounds can be divided in two major areas:

- *Phonology:* studies the significant in the langue, the phonic elements function inside the linguistic communication system.

- *Phonetics:* studies the significant in the parole, the phonic elements production, acoustic constitution and perception.

Debates about the division of phonics into these areas and treating them as separated was made, but years ago they started to being considered as dependent [16].

2.1.4 *Phoneme*

The phoneme is the smallest linguistic unit, that has no own definition but when combined with others, creates a meaning and when substituted by another can change the meaning. For example the words *beso* and *peso* differ in the phonemes */b/* and */p/*, which not only have distinct acoustic properties but also change the word meaning.

Phonemes Classification

Phonemes can be divided into two major types [16, 18]:

- *Vowels*: voice emissions that never find any obstacle when they go over the vocal apparatus.
- *Consonants*: voice emissions that have at least one obstacle when they go over the vocal apparatus.

2.1.5 *Production of articulated sound*

Every time a sound is pronounced (phonemes for example), a variety of movements in our body are performed. These, are produced mainly by three different groups of organs: respiratory, phonation and articulation [18].

The *respiratory* event comprehends inspiration (inhalation) and expiration (exhalation), for the articulated sounds the air exhaled is essential. According to [18] these organs are:

- Diaphragm
- Lungs (left and right)
- Bronchi
- Trachea
- Larynx
- Epiglottis

The *phonation* event in general is a perpendicular tube that directs the air to the articulation organs. The latter, defines the phoneme according to the movements they make, the *articulation* is produced when the organs move from one position to another.

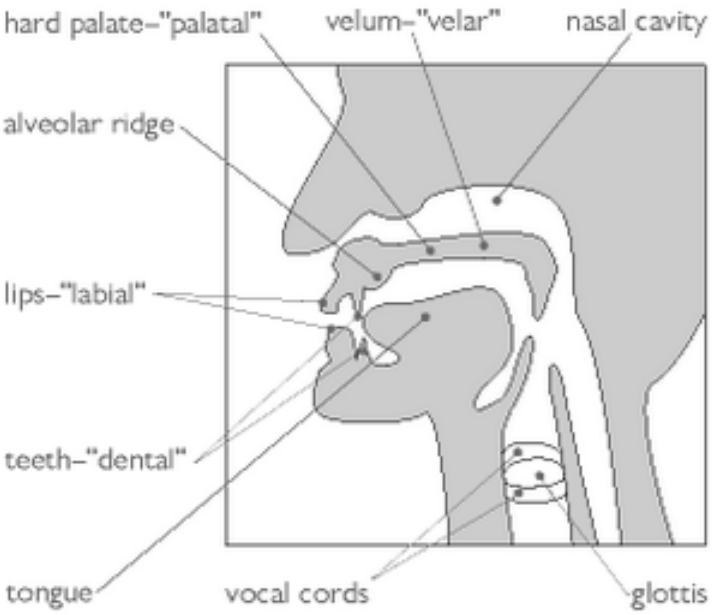


Figure 2.1: Articulation physiological elements

According to [6] the articulation elements are:

- Vocal cords
- Tongue, which makes contact with the:
  - Velum
  - Palate
  - Alveolar ridge
  - Teeth
- Lips
- Nasal cavity

2.1.6 *Articulation point and mode*

The vowel and consonant division for phonemes is too general, that is why is imperative to create subdivisions of these based on two particular aspects:

- *Articulation point*: the place inside the mouth where a certain phoneme is uttered.
- *Articulation mode*: the process that occur when pronouncing a phoneme.

2.1.7 *Vowel classification*

Simple vowels

The Spanish language has five vowels, these are classified as showed in table 2.2. The /a/, /e/ and /o/ are called strong vowels, whereas /i/ and /u/ the weak vowels.

	initials		central		finals
closed	/*i*/				/*u*/
half		/*e*/		/*o*/	
open			/*a*/		

Table 2.2: Simple vowels classification

Vowel triangle

The articulation of vowels can be described according to the tongue position in the oral cavity. As displayed in figure 2.2, a triangle is formed by these five positions. This was introduced by the German Hellwag back in 1781 [18].

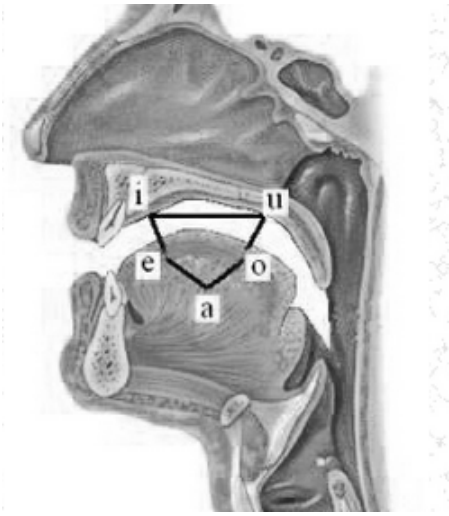


Figure 2.2: Vowel triangle representation

Diphthongs

A diphthong is the combination of a weak vowel and a strong vowel in the same syllable. Therefore they are considered as simple vowels at phonic and orthographic level [6].There is also a classification for them, according to the position of the weak vowel in respect to the other.

	/a/	**/e/**	**/o/**	**/i/**	**/u/**
crescent	/ja/ /wa/	/je/ /we/	/jo/ /wo/		
decrescent	/ai/ /au/	/ei/ /eu/	/oi/ /ou/		
homogeneous				/wi/	/ju/

Table 2.3: Diphthongs classification

2.1.8 *Consonants classification*

The consonants as we described before, are sounds that find at least one obstacle when emitted through the oral cavity. As well, these can be classified by the way they are realized, more specifically by the point of articulation and the mode of articulation.

Point of articulation - classification criteria

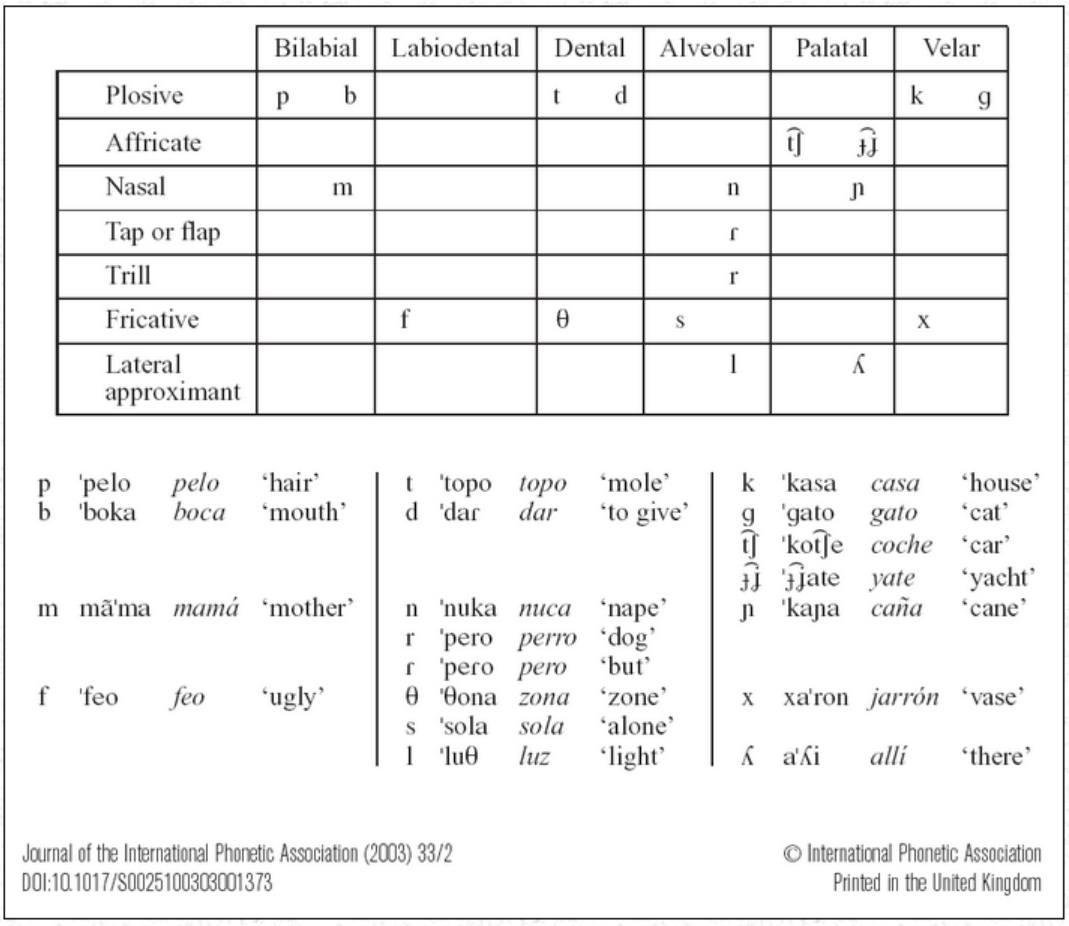
- **Bilabial:** both lips participate.
- **Labiodental:** lower lip with upper teeth.
- **Dental**
  - *Linguointerdental:* the tongue between the teeth.
  - *Linguodental:* the tongue touches the upper teeth rear.
- **Alveolar:** the tongue touches the upper alveoli.
- **Palatal:** the tongue touches the palate.
- **Velar:** the tongue touches the velum.

Mode of articulation - classification criteria

The mode of articulation divides the consonants as followed:

- **Occlusive(plosive):** a explosion is produced.
- **Affricate:** a explosion and rubbing is produced.
- **Nasal:** part of the air is expelled through the nose.
- **Vibrant:** sounds are produced by the tongue vibrations.
  - *Tap or flap:* soundless vibration.
  - *Trill:* with sound vibration.
  - **Fricative:** a rubbing is produced.
  - **Laterals:** the tongue makes the sounds go out through the mouth sides.

The figure 2.3 extracted from the IPA Journal 2003, shows how all consonants are classified in both point and mode of articulation, also some examples are carried out to get a better understanding of their pronunciation.



	Bilabial	Labiodental	Dental	Alveolar	Palatal	Velar
Plosive	p b		t d			k g
Affricate					t͡ʃ ʄ͡ʝ	
Nasal	m			n	ɲ	
Tap or flap				ɾ		
Trill				r		
Fricative		f	θ	s		x
Lateral approximant				l	ʎ	

p	'pelo	*pelo*	'hair'	t	'topo	*topo*	'mole'	k	'kasa	*casa*	'house'
b	'boka	*boca*	'mouth'	d	'dar	*dar*	'to give'	g	'gato	*gato*	'cat'
								t͡ʃ	'kot͡ʃe	*coche*	'car'
								ʄ͡ʝ	'ʄ͡ʝate	*yate*	'yacht'
m	mã'ma	*mamá*	'mother'	n	'nuka	*nuca*	'nape'	ɲ	'kaɲa	*caña*	'cane'
				r	'pero	*perro*	'dog'				
				ɾ	'pero	*pero*	'but'				
f	'feo	*feo*	'ugly'	θ	'θona	*zona*	'zone'	x	xa'ron	*jarrón*	'vase'
				s	'sola	*sola*	'alone'				
				l	'luθ	*luz*	'light'	ʎ	a'ʎi	*allí*	'there'

Journal of the International Phonetic Association (2003) 33/2
DOI:10.1017/S0025100303001373

Figure 2.3: Consonant Phonemes Classification

2.2 Pronunciation Assessment

The teaching of pronunciation comprehends several components and features that can be measured, making it difficult to establish an evaluation framework. As stated out in [21] when it comes to teaching pronunciation is necessary to understand the difficulties that can be encountered:

- The teaching of pronunciation requires that a teacher is devoted into a one to one session with a student, which is almost impossible to achieve in a normal classroom.
- Pronunciation learning demands constant practice requiring a lot of time and patience from the teacher.
- Pronunciation practice can be an exhausting activity due to the use, coordination and control of multiple muscles.
- Normally students prefer not to talk in front of others.
- Testing and grading students pronunciation relies on the teacher subjectivity.

Looking to all these difficulties, teaching pronunciation systems have become more attractive to pronunciation language learners for their portability, availability and cheaper cost. However, it is important to have in mind that the design of these systems is not an easy task, and that a widespread research needs to be done.

2.2.1 *Classification of Pronunciation Errors*

The pronunciation correction of a foreign language (L2) should start by checking previously the errors classification and nature arising in the interlanguage of students. From a linguistic, communicative and foreign accent point of view there are two types of typologies for errors classification: linguistic and communicative [2].

Linguistic Typologies

The linguistic typologies refer to the contrastive analysis over the phonetic-phonological systems of L1 and L2. The error is obtained by detecting the differences over the phonemes distribution among both idioms [2]. Based on this, the errors are:

- *Phonemic errors*: differences among all the phonemes in L1 and L2.
- *Phonetic errors*: interlinguistic equivalence between two similar elements with distinct phonetic statement on the phonological systems of L1 and L2.
- *Allophonic errors*: interlinguistic equivalence among the different allophonic realization of a common phoneme to L1 and L2.
- *Distributional errors*: distinct segment distributions in both L1 and L2.

Communicative Typologies

Communicative typologies on the other hand classify the errors based on the impact caused in the communication, there exist three groups:

- Errors that inhibit communication, like for example confusion over minimal pairs.
- Errors that obstruct communication such as the equivocal utterance of a certain phoneme or group of phonemes.
- Errors that don't difficult communication, these are related to the foreign accent and are considered hard to correct.

Phonemic and Prosodic errors

The quality of pronunciation can be determined by its phonemic and prosodic features [21, 22]. The phonemic most common errors are substitution, deletion or insertion of a phoneme by another. However minor errors can occur such as not very good spoken phoneme, that although, is not all incorrect is different from the native pronunciation of the same phoneme. On the other side the prosodic errors are more concerned with the

stress, rhythm and intonation in a utterance.

Below we list the classification proposed by [22] for the pronunciation error types.

- ***Phonemic***
  - Phoneme Mispronunciation
  - Phoneme Insertion
  - Phoneme Substitution
  - Phoneme Deletion
  - Syllable-level coarticulation errors
- ***Prosodic***
  - Stress
  - Rhythm
  - Intonation

2.2.2 *Computer Assisted Pronunciation Training*

The Computer-Assisted Pronunciation Training (CAPT) has grown in the last years due to the necessity of L2 learners at improving their pronunciation using an automated system. By using CAPT, students can benefit from continuous feedback without a teacher by their side all the time, providing a self-service way to practice [24].

The CAPT field growth has been almost parallel to the evolution of technologies, since computers and mobile devices computing capacity and portability has greatly increased. As pointed out in [22] the CAPT commercialization and work started in the earliest 2000's but it was not until 2007 when its relevance showed up again. With this, the SLATE (Speech and Language Technology for Education) group was created. This group of research is dedicated to the development of education applications by means of automatic speech processing, natural language processing and in some cases spoken dialogue processing. These systems (SLATE/CAPT) make necessary multidisciplinary groups of researchers, from spoken language technologists, language teachers and experts, statisticians, among others [4].

According to several authors the error detection and teaching of pronunciation is a very hard job for CAPT systems, researchers major concern is to derive systems that are capable of identifying errors accurately and reliably to provide correct feedback [4, 19, 22, 24].

2.2.3 *Automatic Speech Recognition in CAPT*

The Automatic Speech Recognition (ASR) often refers to technologies used in the detection and assessment of pronunciation errors, perception training, etc. [4]. The use of ASR

started in the 1980s, but it was not useful for all speakers due to the acoustic differences among them and other related aspects. Then, ASR emerged in actual systems at the beginning of 2000s parallel to the advancements in computing technologies.

The evolution of ASR technologies during the last years has left encountered opinions among researchers about whether or not they are suitable for CAPT systems. In [13] is pointed out that probably the problems found in research are not due to mere ASR but also to the lack of familiarity with the ASR-based CAPT.

Therefore, first we address the phases in which an ASR-based CAPT system can be divided [13].

- *Speech Recognition*: over this phase the incoming speech signal is converted into a sequence of words based on internal phonetic and syntactic models.
- *Scoring*: gives a global evaluation of pronunciation quality by giving a score. Usually the score is obtained by means of comparing temporal properties with the references (natives). Often, this is also called *pronunciation assessment* dealing with the overall impression of fluent speech [4].
- *Error detection*: locating the errors made by the speaker in a utterance, calculating a score and telling them to the speaker. In this case there are two local scores the phoneme in the phonetics approach and syllable or word in prosodics [4].
- *Error diagnosis*: once the error is located, the system is able to specify the speaker the type of error and consequently advise him. These capabilities are the most complex in a system since its necessary to count with particular models of typical errors made by non-natives.
- *Feedback presentation*: this phase comprehends the design issues related on how to present results to the speaker so that in deed help him improving, and also makes the appropriate decisions based on the results presented through the last three phases.

Characteristics

The phases described before give a general point of view on what is necessary for ASR-based CAPT systems, however each of them comprehends diverse aspects to be considered by researchers and that often cause major complications. Obtaining an accurate recognition of the speech concerns to the ASR designers, if the system is not capable of achieving this, then probably teachers and students won't use the CAPT system.

Another fact, is that ASR needs to be adapted for non-native speakers, ASRs developed specifically with native speech have demonstrated worst performance when tested with non-native [13]. To overcome this issue, usually the ASR engine is trained with both native and non-native speech. Knowing the users language (L1) is a must, since the problems that arise when learning a second language (L2) are different in every case and for that matter the ASR needs to understand these before providing feedback [4].

The recognition task also depends on the types of learning activities. Results won't have the same accuracy if there is a huge set of possible answers than when is limited to a small number. Also evaluating an ASR system scoring phase can be done by comparing the scores provided by the human judges for a given speech sample.

The use of a correct combination of scores can help to deliver an accurate error detection to identify the probable pronunciation errors.

2.3 Pronunciation error detection

The pronunciation errors cannot be independently evaluated in a CAPT system, since these are related among them [9, 22]. As we described before, an ASR system is composed by distinct phases, for this case the error detection allows to determine where exactly the user is making a pronunciation mistake. Without an accurate error detection the last phases are useless. Knowing which segment is incorrect and having a database with corrective information of every segment can make possible corrective feedback [4]. Therefore, it is important to understand which features have been proposed to evaluate the pronunciation errors.

2.3.1 *Pronunciation Features*

The pronunciation features for every speech unit (phoneme, syllable, word) help to evaluate how distanced is the non-natives pronunciation from the natives pronunciation. There is an enormous amount of metrics used for evaluating pronunciation. They can be divided by two categories, phonemic and prosodic, since they evaluate the type of errors described before. Table 2.4 depicts the common features and its classification [22].

2.3.2 *Error detection challenges*

The previous section described the features that could be explicitly (although not directly) measured to obtain a possible grading and assessment of a users pronunciation, nevertheless researchers have found more challenges that are significant for the systems performance. According to [22], there are seven core challenges:

- **Reliable phoneme-level error detection**

  The CAPT systems users are aware that mistakes are possible in the pronunciation detection, since even the human experts tend to make inaccurate evaluations [1, 9]. Nevertheless, system designers focus on reducing the *false positive* and *false negatives*. A system is more reliable if these rates are reduced, but more importantly, false negatives need to be kept as low as possible since these directly affect the students performance by decreasing their confidence [4].

- **Distortion error assessment**

  The detection of accent has been little studied, obtaining the degree of accent of the user is difficult since distortions are not easily quantified and classified, an is even more difficult when the L1 and L2 accents are alike.

Feature Category	Feature Name
Phonemic	Phone-level log-likelihood score, GOP Vowel durations, duration trigrams Phoneme Pair classifiers Spectral features (formants) Articulatory-acoustic features
Prosodic (intonation, stress, fluency)	Distances between stressed and unstressed syllables Mean, max, min power per word (energy) F0 contours (slope and maximum) Rate of speech (words per second/minute) Trigram models to model phoneme duration in context Phonation/time ratio, mean phoneme duration Articulation Rate (phonemes/sec) Mean and standard deviation of long silence duration Silences per second Frequency of disfluencies (pauses, fillers etc) Total and mean pause time (i.e. duration of interword pauses)

Table 2.4: Pronunciation Scoring. Features Classification

- **Text independence**

  The use of text reading is often made by pronunciation teachers, which is better for a system since it can be trained with these texts and therefore be more accurate. However this limits the use of spontaneous speaking activities for learning. Anyhow, system designers make use of improved acoustic models of the non-natives and forced-alignment pronunciation evaluation.

- **L1 independence**

  The L1 independence is important from the economic point of view, since to have annotated databases of L1 is expensive and does not scale with the system performance. One way to overcome this is having detailed likely errors just by knowing the users native language.

- **Integrated assessment of both phonemic and prosodic pronunciation components**

  The use of an integrated assessment gives the users more information of their performance. Although, pronunciation error detection has been more into the segmental errors and suprasegmental features, researchers have found that prosody also gives more valuable feedback.

- **Corrective audiovisual feedback**

  There is a lack of suitable software to provide audiovisual feedback to users. Some of them have used cross-sections of the vocal tract to demonstrate the appropriate movements to learners.

- **Robust, interactive system design**

  A variety of approaches to overcome the CAPT issues has been made the past years. In spite of the fact that each approach has its own issues and benefits, efforts are still necessary to integrate them adequately so that CAPT systems are able of providing an end to end series of exercises and activities for pronunciation learning.

2.3.3 *Individual Error Detection*

The pronunciation assessment has often tend for the scoring or pronunciation grading paying less attention to the error detection area. Although error detection comprehends the calculation of a local score, and therefore we would think that error detection is part of scoring because the latter could use local scores to calculate a global scoring, this is not all true, according to [17], both areas have different goals and outcomes.

The error detection is necessary for pronunciation training since it makes possible to give detailed information to the user of where mistakes occur. For example at phoneme level each realization of a phoneme is graded [17]. Obviously, performing these types of measurements depends on different aspects and challenges.

There are multiple approaches for error detection, normally they make use of comparisons, and differences rely on what is being used as reference (non-native or native L2 speech or both) and the type of speech recognition, like forced alignment or unconstrained speech recognition [4].

Techniques like non-native speech adaptation, phonetic processing, confidence measures and other speech and signal processing have been used to accomplish the error detection task [4]. For error detection at phoneme level confidence measures are usually the most used, these can be obtained with the ASR system utilizing Hidden Markov Models [17].

2.3.4 *Confidence Measures*

The confidence measures help determine the certainty of the recognizer when identifying if a utterance (a part or all) was pronounced properly. Confidence measures can be computed with low difficulty by the ASR engine and does not vary greatly among different sounds [17].

A confidence measure can be considered as a statistic that quantifies the fitting of a model with the corresponding data. For speech recognition the acoustic and language models are typically used (together or separately) to extract these confidence measures [20].

The credibility of these measures is obtained by comparing score results with those realized by human judges. In some cases results correlate well enough [24], while others show deficiencies specially when its applied for individual error detection [14].

Likelihood Ratios

Likelihood ratios convert to a useful statistic the outcome of HMM-based ASRs. The HMMs help finding the value of *H*, that maximizes the joint probability *P(X,H)*, where *H* is the acoustic model and *X* is the acoustic observation [20].

$$P(H|X) = \frac{P(X,H)}{P(X)} = \frac{P(X|H)P(H)}{P(X)} \tag{2.1}$$

The Bayes theorem (equation 2.1) is well used to getting the relation of the joint probability with the posterior probability of the model H given the acoustics X, $P(H|X)$ and also the likelihood given the model H, $P(X|H)$.

2.3.5 *Likelihood-based scoring*

There are different proposals that have showed good and well accepted results for automatic scoring, next some of them extracted from the different previous work are listed, considering their results and continuous application.

- **HMM-based Log-likelihood scores**
  The use of likelihood-based phoneme level error detection started back in the 1990s [22]. The log-likelihood logarithm of the speech data is computed by Viterbi algorithm using the HMMs from native speakers. According to [5, 14] this is a good manner for measuring the similarity or matching between native speech and users speech.

  The log likelihood score $\hat{l}$ for each phone segment is [9]:

$$\hat{l} = \frac{1}{d} \sum_{t=t_0}^{t_0+d-1} \log p(y_t|q_i) \tag{2.2}$$

  - $p(y_t|q_i)$:likelihood of the current frame
  - y_t: vector of observations
  - d: duration in frames of the phone segment
  - t_0: starting frame index of the phone segment

  Dividing over d eliminates the time duration of the phone, this way the score is normalized. Also according to [5] the likelihood-based score for a whole sentence L is the average of the individual scores.

$$L = \frac{1}{N} \sum_{i=1}^{N} \hat{l}_i \tag{2.3}$$

  - N: number of phones in the sentence

- **HMM-based Log-posterior Probability Scores**

  The log-posterior probability has showed better correlation results with the human judgments[9].

$$P(q_i|y_t) = \frac{p(y_t|q_i)P(q_i)}{\sum_{j=1}^{J} p(y_t|q_j)P(q_j)} \tag{2.4}$$

  - $P(q_i|y_t)$: is the frame based posterior of the *i-th* phone given the observation vector y_t.
  - $P(q_i)$ represents the prior probability of the phone class q_i.

  The sum over *j* operates on a set of context-independent models for all phone classes. The posterior score $\hat{\rho}$ for the *i-th* segment is the average of the logarithm over the frame-based phone posterior probability over all the frames of the segment, see equation 2.5.

$$\hat{\rho}_i = \frac{1}{d} \sum_{t=t_0}^{t_0+d-1} \log P(q_i|y_t) \tag{2.5}$$

  The complete sentence posterior-based score can be obtained with the average of all individual scores over the N phone segments in the sentence.

$$\rho = \frac{1}{N} \sum_{i=1}^{N} \hat{\rho}_i \tag{2.6}$$

  Compared to the likelihood metric in equation 2.2 the log-posterior probability score could be less affected since the acoustic matching to the models is in both numerator and denominator [5].

- **Segment Duration Scores**

  Evaluating the phone duration is another possibility to obtain a scoring that correlates well with the human expert listener's score based on psychological and linguistic characteristics of the speaker [14]. These characteristics could be:

  - Cross-language differences between L1 and L2
  - Differences in letter-to-sound could provoke insertions, deletions or substitutions of phones which therefore could provoke changes in the segments durations.

  The computing of these segment duration scores comprehends a procedure listed next:

  - Measure the phone duration in frames from the Viterbi alignment
  - Normalize previous results to compensate rate of speech

- Compute each phone-segment duration score using the log-probability of the normalized duration.
- The latter log probabilities are obtained by computing discrete distributions of duration for the respective phone.
- These distributions are trained with alignments generated from the native training data.

$$D = \frac{1}{N}\sum_{i=1}^{N}\log(p(f(d_i)|q_i) \tag{2.7}$$

- $f(d_i)$: duration normalization function
- q_i: phone q that corresponds to the ith segment

To accomplish text independence the rate of speech (ROS) is utilized as a normalization factor, according to [14, 15],

2.4 Goodness of Pronunciation (GOP)

The last section depicted some of the most important likelihood-based scores that have been utilized for automatic scoring of the phoneme pronunciation. According to [8, 4], another best example based on likelihood is the Goodness of Pronunciation (GOP) algorithm proposed by Witt [23, 25], which relies on confidence scores derived from recognition results. Next, we do a special description of this metric and we look over some implementations along with the variants that have been realized in some of the state of the art works.

2.4.1 *The Basic GOP*

The GOP was first introduced by Witt with the purpose of providing an algorithm capable of scoring each phone of an utterance, therefore to accomplish this, the GOP must have previously the following data [23]:

- The *orthographic transcriptions* previously annotated by human judges that describes exactly which is the phone sequence uttered.
- The *Hidden Markov models* to calculate the *likelihood*, $p(O^{(q)}|q_j)$, where $O^{(q)}$ is the acoustic segment corresponding to each phone q_j.

Based on the latter the GOP for a phone q_i is computed by:

$$GOP_1(q_i) = |\log(P(q_i|O))|/NF(O^{q_i}) \tag{2.8}$$

Based on equation 2.8, the quality of pronunciation of any phone q_i can be obtained by normalizing the logarithm $P(q_i|O^{q_i})$, which is the posterior probability that the speaker uttered the phone q_i over the acoustic segment O^{q_i}. The normalization takes place when dividing by the number of frames ($NF(O^{q_i})$) in the acoustic segment.

As showed in equation 2.4 the log-posterior probability can be computed by knowing the likelihood of the acoustic observations given the phone q_i and the likelihood of the acoustic observations given the phone models. By applying this we get:

$$GOP_1(q_i) = |\log\left(\frac{p(O^{q_i}|q_i)P(q_i)}{\sum_{j=1}^{J} p(O^{q_i}|q_j)P(q_j)}\right)|/NF(O^{q_i}) \tag{2.9}$$

For equation 2.9 J is the total number of phone models for the j possible phones existent in the annotated database. If we assume that all phones are equally likely, meaning $P(q_i) = P(q_j)$ for all j and i, and that the sum of the denominator can be approximated by its maximum, then the GOP is equal to:

$$GOP_1(q_i) = |\log\left(\frac{p(O^{q_i}|q_i)}{\max_{j=1}^{J} p(O^{q_i}|q_j)}\right)|/NF(O^{q_i}) \tag{2.10}$$

The numerator of equation 2.10 is computed by using the forced alignment block showed in figure 2.5, in this the sequence of phone models is fixed by the known transcription. On the other hand, the denominator is obtained by the Phoneme Loop block, which realizes an unconstrained loop comparing the acoustic observations of the i-th phone with all the possible phonemes transcription.

The GOP in equation 2.10 can be rewritten as:

$$GOP_1(q_i) = \left|\frac{\log(p(O^{q_i}|q_i))}{NF(O^{q_i})} - \frac{\max_{j=1}^{J}\log(p(O^{q_i}|q_j))}{NF(O^{q_i})}\right| \tag{2.11}$$

When mispronunciations occur, the alignments from the phone loop will not match those from the forced alignments as showed in figure 2.4. The latter means that more than one phone in the unconstrained phone sequence affects the calculation of $\max_{j=1}^{J} p(O^{q_i}|q_j)$. Since N phones could contribute to the phone loop computation, then the calculation takes place as follow:

$$\frac{\log(p(O^{q_i}|q_j))}{NF(O^{q_i})} = \sum_{i=1}^{N} \frac{\log(p(O^{q_i}|q_{ji}))}{f_{ie} - f_{is}} \tag{2.12}$$

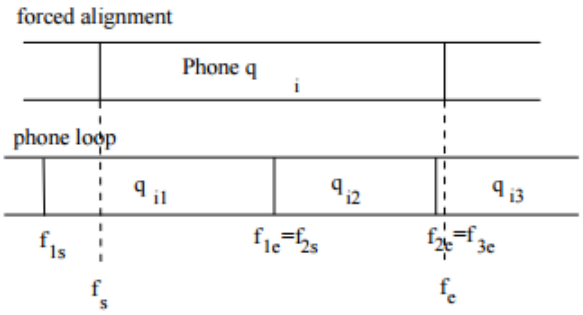


Figure 2.4: Comparing the phone loop alignments with the forced alignments when a mispronunciation occurs[24].

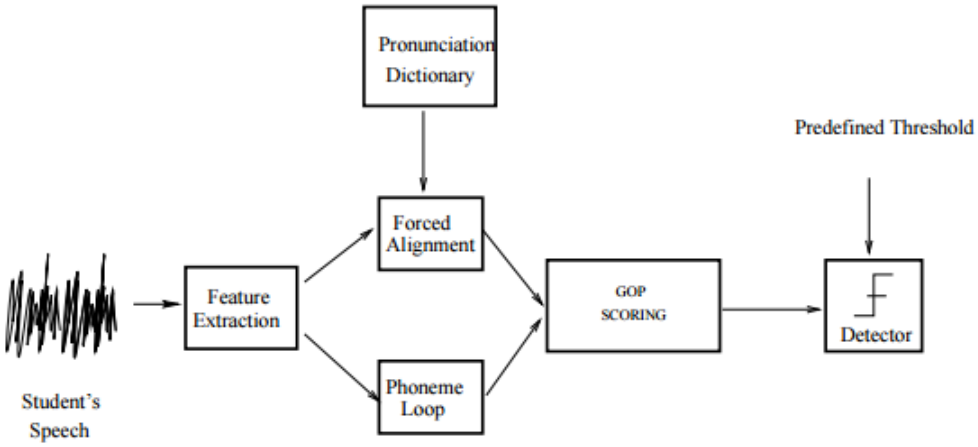


Figure 2.5: Block-diagram of a pronunciation scoring system [24].

The figure 2.5 depicts the computational phases inside the scoring system proposed by Witt [25].

- *Feature extraction*: the extraction of a framed representation of the speech wave with their corresponding mel-frequency cepstral coefficients (MFCC) takes place here.
- *Forced alignment*: fixes the sequence of phone models with the known annotated transcription.
- *Phoneme Loop*: an unconstrained loop looks for the most likely phone model that fits the acoustic observation over all the frames.
- *GOP scoring*: computes the formulas exposed before.

- *Thresholds*: applying a threshold to each GOP score to reject badly pronounced phones. These thresholds vary according the correctness desired.

2.4.2 *An overview over GOP variants*

The implementation of the GOP metric has been widely used by several groups of researchers with different native and non-native languages. Next, some of these are described along with some of the most important characteristics and variants that have been added to improve the results performance.

Thresholds dependency

In the thesis of Witt [25] some thresholds calculation formulas are proposed to improve results. These bring forward the use of phone dependent thresholds instead of a single phone threshold for all phones due to that acoustic fit of phone based HMMs differs from phone to phone.

One proposal for calculating a phone-specific threshold is computing the global GOP statistics. A specific threshold for a phone q_i would be:

$$T_{q_i1} = \mu_{q_i} + \alpha\sigma_{q_i} + \beta \tag{2.13}$$

Where:

- μ_{q_i} is the mean of all the GOP scores of the phone q_i
- σ_{q_i} is the variance of all the GOP scores of the phone q_i
- α and β are empirical constants

The second proposal makes use of the human labeling behavior to obtain a second threshold.

$$T_{q_i2} = \log \frac{1}{N} \sum_{n=1}^{N} \left(c_n(q_i) \Big/ \sum_{m=1}^{M} c_n(m) \right) \tag{2.14}$$

Where:

- $c_n(q_i)$ is the total number of times that phone q_i was labeled as mispronounced by a human judge.
- N is the total number of speakers
- M is the total number of phones

Explicit error modeling

According to [24] pronunciation errors can be classified in two different error classes:

- *Individual mispronunciations*: the pronunciation of a specific word is unknown for the student.
- *Systematic mispronunciations*: substitutions of native sounds for sounds of the target language, which do not exist in the native language.

The GOP formula studied before doesn't take into account students native language phone modeling. Therefore, if systematic mispronunciations are realized, then the acoustic modeling of the non-native speech will be incorrect.

To overcome this issue Witt [24] used a recognition network that comprehended both correct pronunciation and common pronunciation errors as sublattices for each phone. These were based on phone models of target and native language.

The recognition network derives a sequence of phones that in the best case will likely suit the target transcription, meaning $q_i = q_{it}$ or on the contrary an a error phone will be obtained, $q_i = q_{ie}$. In case a error is detected by the recognition network, is also necessary to take into notice the likelihood of its occurrence and not just make a yes or no decision.

The latter leads to get the posterior likelihood of the phone error, $P(q_{ie}|O^{(q_i)})$. It can be computed with the equation 2.10, where the acoustic phone models of L1 and L2 will be used in the phoneme loop calculations.

Another fact is that by acquiring the posterior likelihood of q_{ie} is possible to calculate the posterior probability of the target phones q_{it}, for all the phones with systematic mispronunciations.

$$P(q_{it}|O^{(q_i)}) = 1 - \sum_{q_j \neq q_{it}} P(q_j|O^{(q_i)}) \tag{2.15}$$

If only the maximum (assuming the sum can be approximated to it) of all the phone probabilities distinct from q_{it} is considered then equation 2.15 is:

$$P(q_{it}|O^{(q_i)}) \approx 1 - \max_{q_j \neq q_{it}} P(q_j|O^{(q_i)}) \tag{2.16}$$

The probability of a given phoneme $q_j \neq q_{it}$ is maximum when $q_j = q_{ie}$, thus we get:

$$P(q_{it}|O^{(q_i)}) = 1 - P(q_{ie}|O^{(q_i)}) \tag{2.17}$$

Then, the new scoring $GOP_e(q_i)$ is defined as:

$$GOP_e(q_i) = \begin{cases} |\log(1 - P(q_{ie}|O^{(q_i)})| & \text{if } q_i = q_{ie}, \\ 0.0 & \text{otherwise.} \end{cases} \tag{2.18}$$

The combination of equations 2.10 and 2.18 produces a new GOP metric that now considers the occurrence of systematic errors.

$$GOP_2(q_i) = GOP_1(q_i) + KGOP_e(q_i) \tag{2.19}$$

Where K is a scaling constant.

Other implementations

Apart from the variants discussed before, many other researchers have used the GOP metric in their works and studies. The adding of new features have in some cases turn out in better results. Basically these consist on techniques to establish new and more accurate thresholds, more solid and better annotated corpus for training and obtaining the acoustic models and some other optimization criteria.

The PLASER system is a multimedia tool with instant feedback designed to teach English pronunciation to students from Hong Kong whose native language is Cantonese Chinese [11]. This system uses equation 2.10 to obtain the GOP metric of the phone acoustic segment being evaluated. Also, the phoneme loop (denominator in the equation) is replaced by the Viterbi likelihood of the segment given by a phone loop. The GOP results are then passed through thresholds but there is no clear explanation of how they have been established.

Another GOP variant proposed for the PLASER system is **normalizing the GOP score** in a range [0.0 ... 1.0], this is accomplished by using a **sigmoid function**, see equation 2.20.

$$sigmoid(GOP_{q_i}) = \frac{1}{1 + \exp(-\alpha GOP_{q_i} + \beta)} \tag{2.20}$$

Where α and β are empirical constants. Take notice that these represent distinct values than the latter.

Afterwards, the GOP scores are computed to give two different visualization of recognition results:

- *an overall phoneme score of the whole word*
- *a phoneme-by-phoneme assessment by a 3-color scheme*

The **overall phoneme score** (*PS*) for a word is the result of a weighted sum of the normalized phones GOP that compose it.

$$PS(word) = \sum_{k=1}^{N} w_k \cdot normalizedGOP(q_{ik}) \tag{2.21}$$

Where w_k is the weighting of the k-th phoneme among the N phonemes composing the word.

As well the **phoneme-by-phoneme assessment** yields a three color representation for a bad, fair or good pronounced phoneme, that is mapped into the corresponding letter in the word. As mentioned in their article the settlement of these thresholds is made through an algorithm that relies on the best bi-threshold nature, meaning having acceptable values of *false acceptance rate* (FA) for an incorrectly pronounced phoneme and of *false rejection rate* (FR) for a correctly pronounced phoneme.

Likewise, the work realized in [8] proposes the calculation of new thresholds on evaluating pronunciation errors frequently made by foreigners speaking Dutch. Their approach is divided on three experiments that evaluate the GOP:

- using a native test corpus with artificial errors which reflect errors frequently made by non-natives.
- within an actual application used by non-natives for practicing pronunciation
- post-hoc, using the recorded interactions of the pronunciation training application, to determine what the performance of the algorithm would have been if optimal speaker and phone specific thresholds had been used.

The goal was establishing GOP thresholds that maximize the Scoring Accuracy (SA) performance measure and at the same time maintain the False Rejections rate (FR) under 10%. Optimal thresholds were encountered for each phoneme-gender combination, showing that a step size of around 0.25 worked well enough. Later, for the application these new thresholds were applied showing a better SA and finally the threshold for each phoneme-speaker pair that showed the top SA was calculated, then the performance for each speaker and last the combination of all values from the speakers to obtain global measurements for the whole group.

Similarly, the proposal from [17] used the GOP metric focusing on the establishment of thresholds that best suited the data. Although they tested the use of the type 1 threshold discussed before (see equation 2.13) the results where not good. After, they search for thresholds that maximize the SA and the FR under 10% similar to the one evaluated by [8].

The Forced GOP

The GOP implementations depicted so far have focused more on the establishment of distinct thresholds after the GOP computation. However, another interesting variant is the one detailed in [10] and that has also been implemented by our research group to obtain the data analyzed in this work.

This new variant of the GOP utilizes the forced alignment block as in Witt allowing to set the phoneme boundaries. Once these boundaries are known the computation of the $p(O^{q_i}|q_j)$ (logarithmic scores) for all the j existent phonemes is realized.

Thus, to obtain the GOP, is necessary to determine:

- The annotated(orthographic transcription) utterances that will allow to determine which is the expected phoneme q_i.
- All the log likelihoods for every phoneme $p(O^{q_i}|q_j)$.

Afterwards, the computation of the GOP is realized by subtracting the expected phoneme q_i logarithmic score with the maximum logarithmic score among all phonemes, as in formula 2.11.

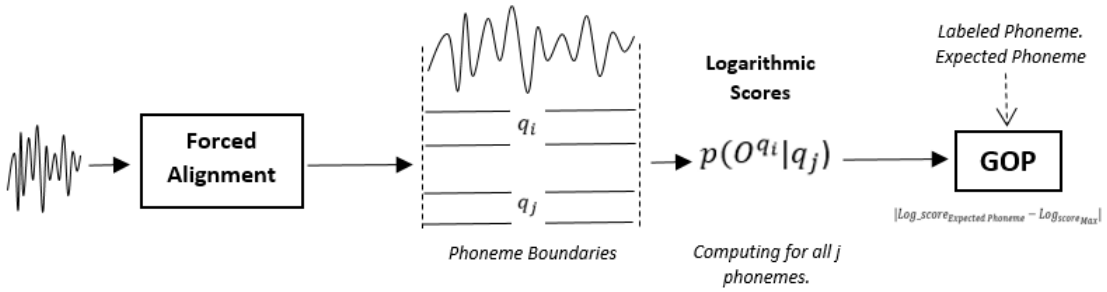


Figure 2.6: Block-diagram of the forced GOP computation

Chapter 3

Tests and Results

3.1 Materials

3.1.1 *Corpus description of the evaluated utterances*

The evaluation of the GOP metric results discussed later has been done over a subset of the SAMPLE corpus. Likewise, the latter is based on a subset of the GLISSANDO corpus, which comprises a big set of sentences and paragraphs derived from the news database of a popular Spanish radio news [7]. These depict a diversity of topics closely related to daily life .

The SAMPLE corpus is composed of the uttered sentences and paragraphs produced by 14 non-native speakers of Spanish studying this language at university, 9 American English and 5 Japanese. Also, 8 native Spanish speakers with different speech features were recorded [1].

Likewise, this work uses a subset of the SAMPLE corpus. This selection includes the 15 sentences uttered by all the 14 non-native speakers and the 8 natives. The non-native speakers uttered the sentences more than one time, in some cases up to 6 repetitions were made and results are equally used depending on the type of measure or testing realized.

Feature	**Description**
Number of Speakers	22 (14 non-natives and 8 natives)
Native Speakers ID	L01, L02, L03, L04, L05, L06, L07, L08
Non-Native Speaker ID	f01, f02, m03, f04, f05, f06, f07 m08, f09, f10, f11, f12, f13, f14
Number of sentences	15
Number of repetitions	Natives: 1, Non-Natives: 1 to 6

Table 3.1: Subset of the SAMPLE corpus features

Along the results each speaker is identified as showed in table 3.3.

S-ID	Sentence
s01	La coalición interpuso esta querella por prevaricación el viernes pasado.
s02	52 denuncias por faltas graves en dos años, 18 de ellas graves por carecer de licencia de funcionamiento. Y el bar sigue abierto.
s03	Para una gala que se celebrará el 8 de febrero del próximo año.
s04	ATT prevé eliminar 12.000 empleos y reducir inversiones de capital.
s05	Notó una foto con flash cuando volvía a su domicilio.
s06	En la cartelera de cine no hay este fin de semana mucha poesía que digamos.
s07	¿Qué sería de una Navidad sin su cesta?
s08	Más de un millón de mujeres trabajan actualmente por cuenta propia.
s09	Y en los mercados los números rojos se extienden hoy por todas las bolsas europeas.
s10	No les han ofrecido hotel, ni tan siquiera a un vaso de agua.
s11	Sin embargo, también hay una buena noticia. existen soluciones.
s12	Sigue con sus trabajos de investigación, en los que ya constan sus conversaciones con la presidenta regional.
s13	Todos ellos, según las últimas informaciones del diario El País, fueron también víctimas de seguimientos.
s14	Esta investigación interna no ha dado aún ningún dato concluyente, y no tiene fecha límite.
s15	Hoy, hay huelga en las escuelas infantiles.

Table 3.2: Set of sentences uttered by the different speakers

Representation	ID
Speaker 1	f01
Speaker 2	f02
Speaker 3	m03
Speaker 4	f04
Speaker 5	f05
Speaker 6	f06
Speaker 7	f07
Speaker 8	m08
Speaker 9	f09
Speaker 10	f10
Speaker 11	f11
Speaker 12	f12
Speaker 13	f13
Speaker 14	f14
Speaker 15	L01
Speaker 16	L02
Speaker 17	L03
Speaker 18	L04
Speaker 19	L05
Speaker 20	L06
Speaker 21	L07
Speaker 22	L08
Speaker 23	Albayzin

Table 3.3: Representation of the speakers along the results display

3.1.2 *Corpus description of the GOP results*

The sentences described before were uttered by every speaker and recorded in a audio studio to evaluate the speakers phoneme pronunciation. The orthographic transcriptions used to identify the uttered phones in every sentence were realized by a group of linguists that collaborated with our research group. Also, the implementation of the forced GOP described in 2.4.2 was done by Cesar González as a internal work for our research project.

The Hidden Markov Models were used to calculate the $p(O^{q_i}|q_j)$. The training of these models is based on a standard parameterization by using cepstral coefficients over mel frequencies (MFCC) and a 39 dimensions feature vector. More precisely, 12 MFCCs and the normalized power logarithm along with the first and second order derived. The features vectors are obtained with a time slot of 25 ms and time offset of 10 ms. The Albayzin corpus was used to train the acoustic mono-phoneme models since this contains recordings of phonetically balanced phrases [12].

Finally, all the sentences uttered by the 22 speakers of the SAMPLE corpus were computed by the Forced GOP implementation. The results present the logarithmic scores (likelihoods) of all existent phonemes for every expected phoneme in a uttered sentence for a given repetition number and the speaker who uttered it grouped into one file. All these files were computed to analyze the GOP and the Logarithmic Scores along this work.

Expected	***a***	
GOP	***0.0***	
Phoneme Logarithmic Scores	a	-82.450127
	b	-97.913742
	d	-98.16037
	e	-87.589706
	f	-114.805275
	g	-94.814545
	i	-103.71199
	j	-101.506226
	J	-106.872292
	jj	-107.73539
	k	-102.665993
	l	-94.937546
	L	-114.148155
	m	-104.059875
	n	-102.17115
	o	-86.874405
	p	-115.116127
	r	-90.437248
	rr	-94.109604
	s	-102.456642
	t	-113.700119
	T	-113.858459
	tS	-117.591797
	u	-96.584976
	w	-94.169151
	x	-103.48201

Table 3.4: GOP and Log scores results for a given phoneme example 1

The table 3.4 is an example of the results obtained for the expected phoneme ***a*** when the GOP value obtained is zero and table 3.5 gives an example of results when GOP is distinct from zero.

The GOP value obtained (second row) for every phoneme is a score of how good it was uttered, the closest this value is to zero means a better pronunciation. As depicted in 2.4.2 the forced GOP computes the log likelihood scores for all j possible phonemes, which in our case are 26. As showed in tables 3.4 and 3.5 each phoneme has a logarithmic score, and the GOP value of the i-th phoneme (expected phoneme) is the absolute value of the subtraction between the maximum logarithmic score and the logarithimic score of $j = i$ phoneme.

The following annotation represents the phonemes as follows, symbol= phoneme respectively: J = *eñe*, T = *ce*, tS = *che* and jj = *ya*. For computing purposes the symbols are changed as follow jj=$, rr=& and tS=#.

Expected	***n***
GOP	***2.315826***
Phoneme	**Logarithmic Scores**
a	-91.01696
b	-85.112442
d	-86.827423
e	-96.432953
f	-106.900955
g	-87.880821
i	-107.076622
j	-101.933418
J	-94.041672
j	-102.059052
k	-97.649231
l	-87.85672
L	-102.924179
m	-88.893578
n	-87.428268
o	-91.635689
p	-107.062897
r	-86.42289
rr	-86.099121
s	-102.498901
t	-108.815567
T	-103.24295
tS	-110.937195
u	-93.996315
w	-90.972496
x	-98.107735

Table 3.5: GOP and Log scores results for a given phoneme example 2

3.2 Results

Along the results section all the experiments and statistical analysis realized over the data is described. First, the evaluation of the GOP is conducted. Second, a statistical review is done over the logarithmic scores of all phonemes derived in every utterance of a expected phoneme. Third, we take a look to the maximum logarithmic score value obtained for the calculation of expected phoneme GOP. Fourth, we go into a deeper level by evaluating the logarithmic score obtained for every expected phoneme in two cases; when the GOP value was zero (success) and when the GOP value was different from zero (failures). Then, we analyze the behavior of a expected phoneme logarithmic score with its corresponding GOP value. Finally, some new scoring proposals are detailed and the different global scorings (including the GOP results) for every speaker are contrasted with the human judgments.

3.2.1 *GOP evaluation over the SAMPLE Corpus*

The evaluation of the GOP metric over the SAMPLE corpus started by computing a global score for every speaker. In this way we would be able to compare the results against the human judgments. To obtain a global score for each speaker the average of all the GOP values of every expected phoneme uttered by the same speaker was calculated, results are showed in table 3.6.

	Speaker	Global FGOP score
Non-Natives	f01	3.52
	f02	3.75
	m03	**5.14**
	f04	3.77
	f05	3.55
	f06	3.75
	f07	4.03
	m08	3.95
	f09	3.98
	f10	4.35
	f11	3.81
	f12	3.00
	f13	**2.96**
	f14	3.17
Natives	L01	3.57
	L02	3.10
	L03	3.64
	L04	**2.72**
	L05	4.46
	L06	**4.46**
	L07	3.84
	L08	4.11
	Albayzin	0.8702

Table 3.6: Global Forced GOP scores for every speaker.

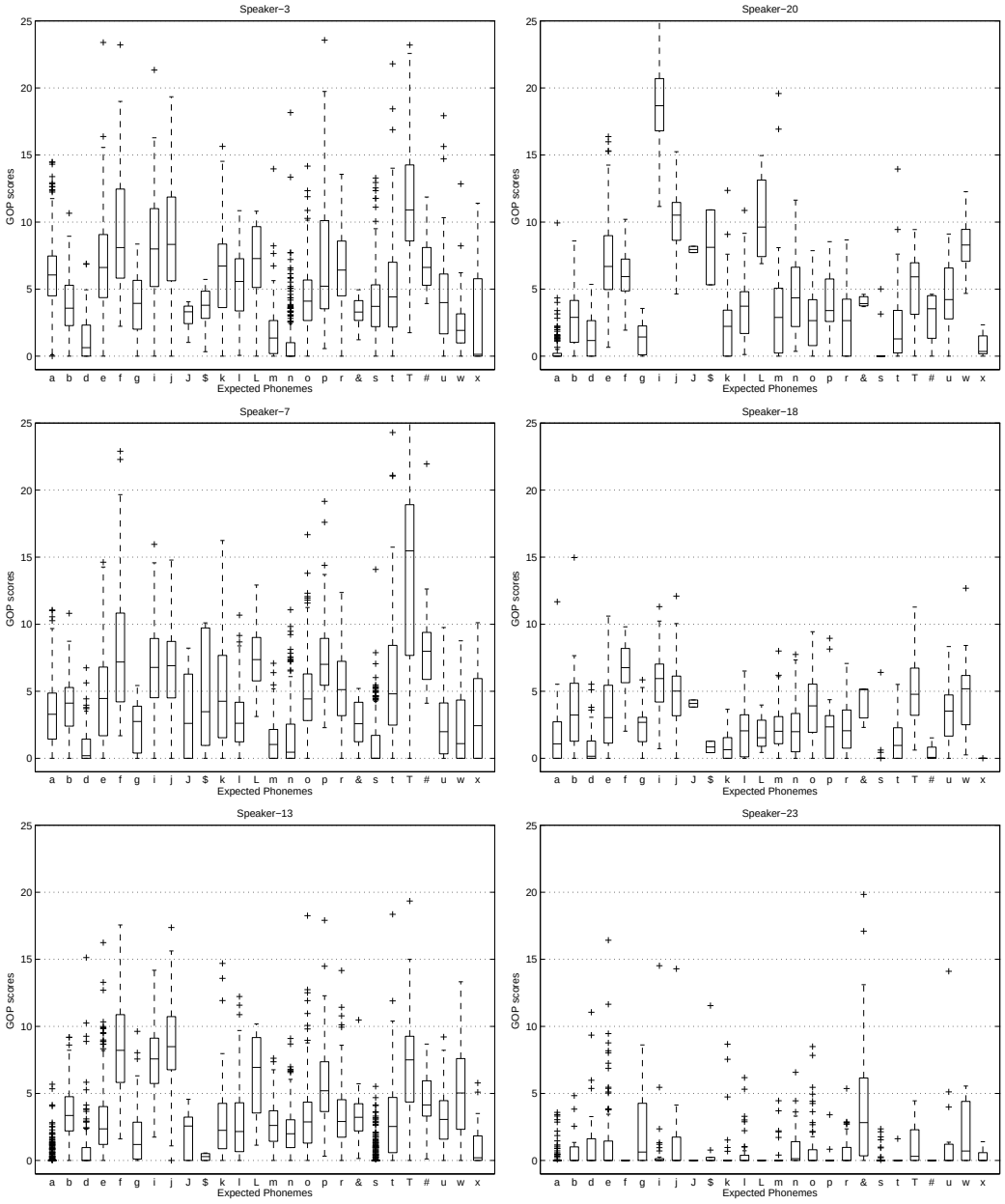


Figure 3.1: **GOP boxplots comparison of the worst to the best speakers.** First column Non-Natives: m03, f07, f13 and second column Natives: L06, L04, Albayzin.

A speaker is better than the others if its GOP score is closer to zero, these results show that the best speaker is Albayzin, which was a expected result since the models for the FGOP were trained with it. The speaker *L04* is the best among natives and non-natives, and the worst speaker is *m03*. Among non-natives the best speaker was *f13*.

The results are not good for natives when compared to the non-natives results, we would have expected that the GOP global score values for natives were lower than the non-native scores.

The figure 3.1 represents the distributions of the GOP values for every expected phoneme of some speakers. The first column shows three non-natives and the second column three natives. The first row has the worst speakers of both groups *m03* and *L06*, the second row shows speakers with intermediate results and the third row has the best speakers of each group *f13* and *Albayzin*.

The graphics show that Albayzin distributions are the closest to zero. The worst speaker 3 (m03), has distributions that barely get near the zero (just phoneme *d* and *m*). Speaker 20 (L06) has more distributions near the zero, however some are also far away from the zero (look phonemes *i* and *j*). The speaker 7 (f07) shows better results, more of its distributions are closer to zero. The best non-native speaker 13 (f13) shows the best distributions not only because they are closer to zero but also have less deviation (smaller boxes).

The results also show that some phonemes are more problematic for the non-natives than for natives, like for example phonemes *T* and *f* have worst distributions for non-natives (even for the best speaker) than for native speakers. Phoneme *a* shows not to be a problem for natives, but for the non-natives (m03 and f07) its distributions are not good.

Figures 5.1, 5.2, 5.3 and 5.4 in Appendix A, show the GOP boxplots for all the speakers.

3.2.2 *Logarithmic Scoring Evaluation*

Logarithmic scores

The logarithmic scores (log-likelihoods) of all phonemes that are computed by the FGOP algorithm are also analyzed. As explained before the GOP value is the subtraction of the maximum logarithmic score and the log score associated to the expected phoneme. So the GOP value depends directly on these values.

The graphics on figure 3.2 represent the distributions of the logarithmic scores for all phonemes in the corpus classified by speaker.

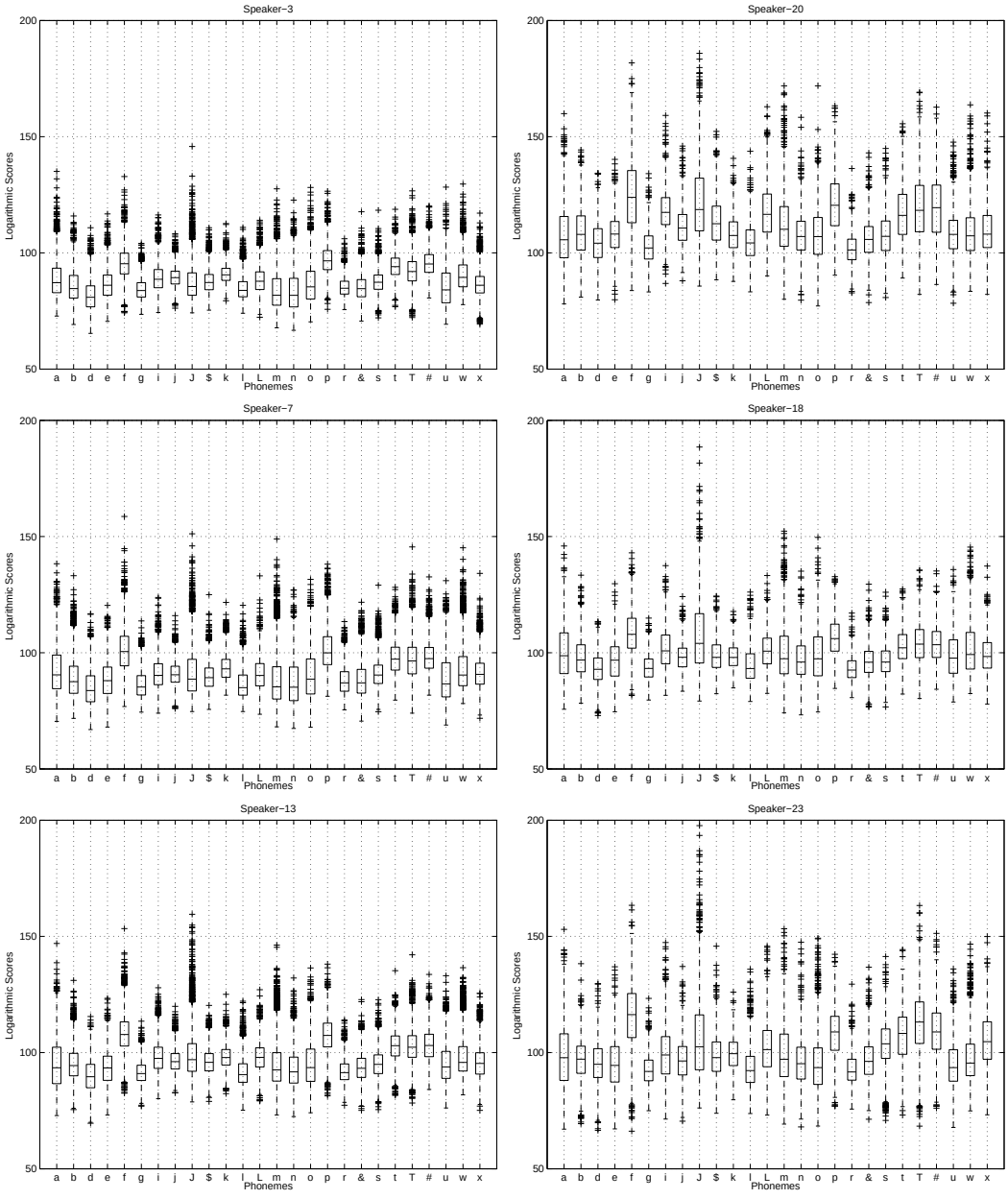


Figure 3.2: **Logarithmic scores** for every phoneme. Boxplots comparison of the worst to the best speakers. First column Non-Natives: m03, f07, f13 and second column Natives: L06, L04, Albayzin.

Figure 3.2 represents the same six speakers analyzed in the section before. The first column has the non-natives m03, f07 and f13, and the second column the native speakers L06, L04 and Albayzin, ordered according its GOP score, from the worst to the best.

The logarithmic scores represent the likelihood of the acoustic segment (uttered phoneme) for every phoneme's model. These models were obtained with the Albayzin corpus, which means that the logarithmic scores distributions obtained for Albayzin are the ones that maximize the best result and therefore are the reference.

The logarithmic scores of every phoneme mark a particular trend for all the speakers which means that there is a phoneme dependency associated to the calculation of the logarithmic score. The best speakers of both groups f13 and L04 depict distributions more alike with Albayzin. The speaker f07 distributions are a little below the average values of Albayzin. Finally and more interesting, the distributions of the worst speakers (row 1) have values that are farther from the values of Albayzin.

The worst non-natives usually have logarithmic scores that are closer to zero than the values from Albayzin. On the other hand, the worst natives have values that are farther from zero than the values from Albayzin.

Figures 5.5, 5.6, 5.7 and 5.8 in Appendix A, show the logarithmic scores boxplots for all the speakers.

Maximum Logarithmic scores

In this section we go a little bit deeper by only looking the maximum logarithmic score obtained when computing the GOP value for every expected phoneme that has been uttered by a determined speaker. The graphics have the absolute values of the logarithmic scores, meaning that the maximum values are the minimum in the plots.

Figure 3.3 represents the same six speakers that we have been analyzing. The first column has the non-natives m03, f07 and f13, and the second column the native speakers L06, L04 and Albayzin, ordered according its GOP score, from the worst to the best.

There are no evident differences among the non-natives distributions, meaning that the maximum logarithmic score has little or no correspondence with the speaker. For the natives on the other hand, the distributions values vary greatly and if we compare each graphic to its corresponding when all logarithmic scores were analyzed the decrease on the average values is almost constant.

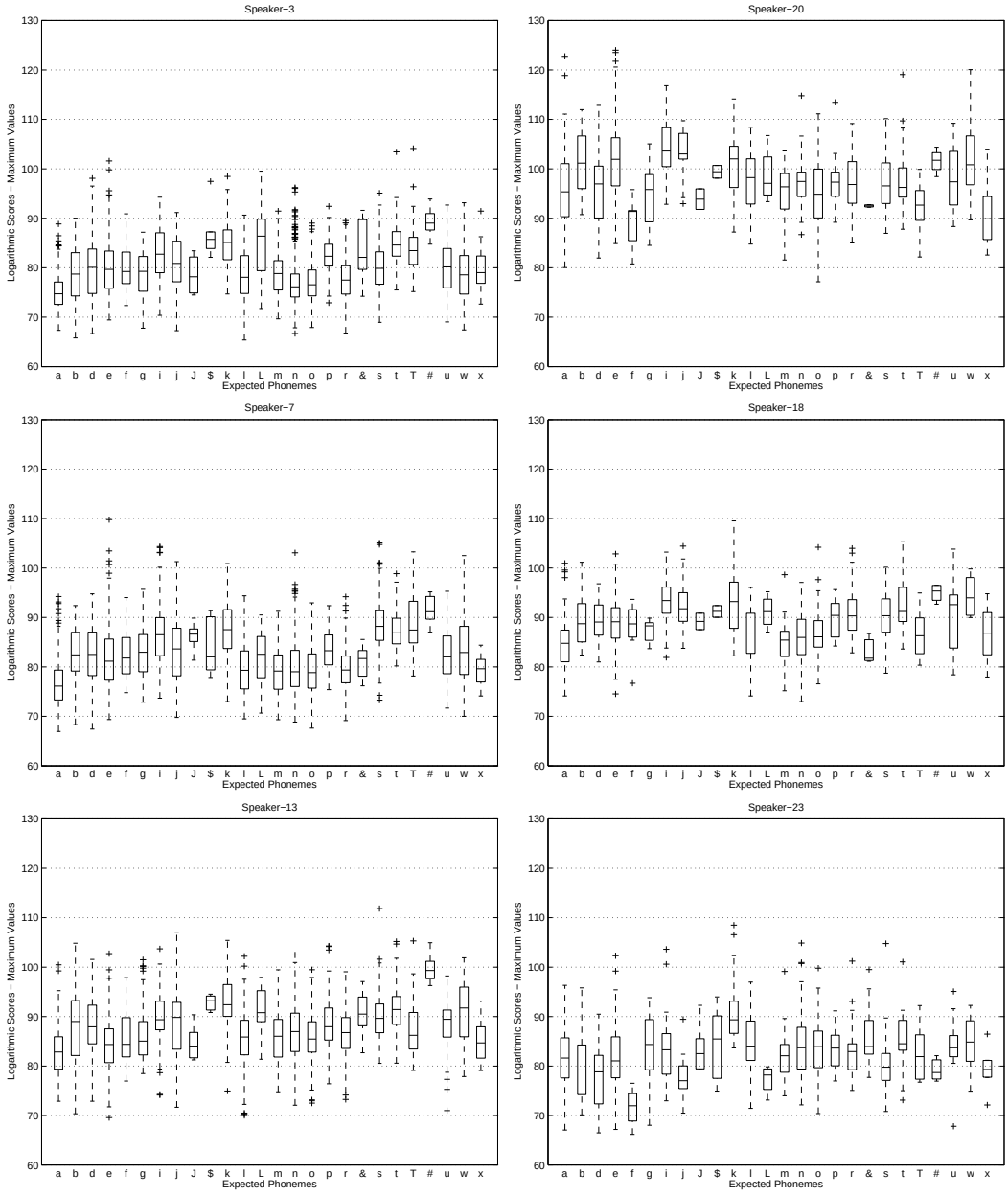


Figure 3.3: **Maximum logarithmic scores** for every phoneme. Boxplots comparison of the worst to the best speakers. First column Non-Natives: m03, f07, f13 and second column Natives: L06, L04, Albayzin.

Success and Failure. Logarithmic scores

To continue with the logarithmic scores evaluation, the logarithmic scores associated to the expected phoneme are analyzed in two situations. First, when the GOP was equal to zero, meaning the maximum logarithmic score was equal to logarithmic score of the expected phoneme, which are called the success cases. Second, when the GOP was different from zero.

Figure 3.4 shows the graphics of the worst speakers (m03 and L06) among both groups of non-natives and natives respectively, and at the bottom the graphics of Albayzin appear. Likewise, figure 3.5 shows the graphics of the best speakers (f13 and L04) among both groups of non-natives and natives respectively and also the graphics of Albayzin in the third row. The graphics over the first column represent the success case and graphics on the second column the failure cases.

The first clear observation is that over the success case the worst speakers have none distributions for the majority of the expected phonemes. The best speakers have more appearances but not as much as Albayzin. The last demonstrates that the speakers with poorest pronunciation quality are the ones that uttered less correct phonemes.

The best speakers show less variation among the success and failure distributions for the same phoneme, and also in general, while the worst speaker show higher variations.

Figures 5.9, 5.10, 5.11 and 5.11 in Appendix A, show the success and failure graphics for all the speakers.

GOP and Logarithmic Scores for groups of phonemes

Finally, we represent for some of the expected phonemes their GOP values against their corresponding logarithmic score (all the utterances in the corpus are used). The figure 3.6 represents the values for the five vowels ordered like in Hellwag's triangle. These show that the closed vowels *i* and *u* were less uttered and also have less occurrences with GOP equal to zero, the vowel *i* has the worst results with GOP values that reach twenty. The open vowel *a* shows the best performance when compared to the others.

Figure 3.7 represents the plosive consonants, *p*, *b*, *t*, *d*, *k* and *g*. These show that phoneme *p* had bad performance, since almost no GOP value was equal to zero, on the other side, phoneme *d* shows good performance, there is a concentration of occurrences near the GOP equal to zero and less occurrence's with high values of GOP.

Finally, figure 3.8 represents the graphics of the alveolar consonants, *n*, *r*, *rr*, *s* and *l*. The phoneme *rr* has very few occurrences and most of them show bad grading (GOP distinct from zero), meaning that this is a problematic phoneme for the speakers. For all the other phonemes, although they have high concentration around the GOP equal to zero they have also many bad results and no specific distinction among them.

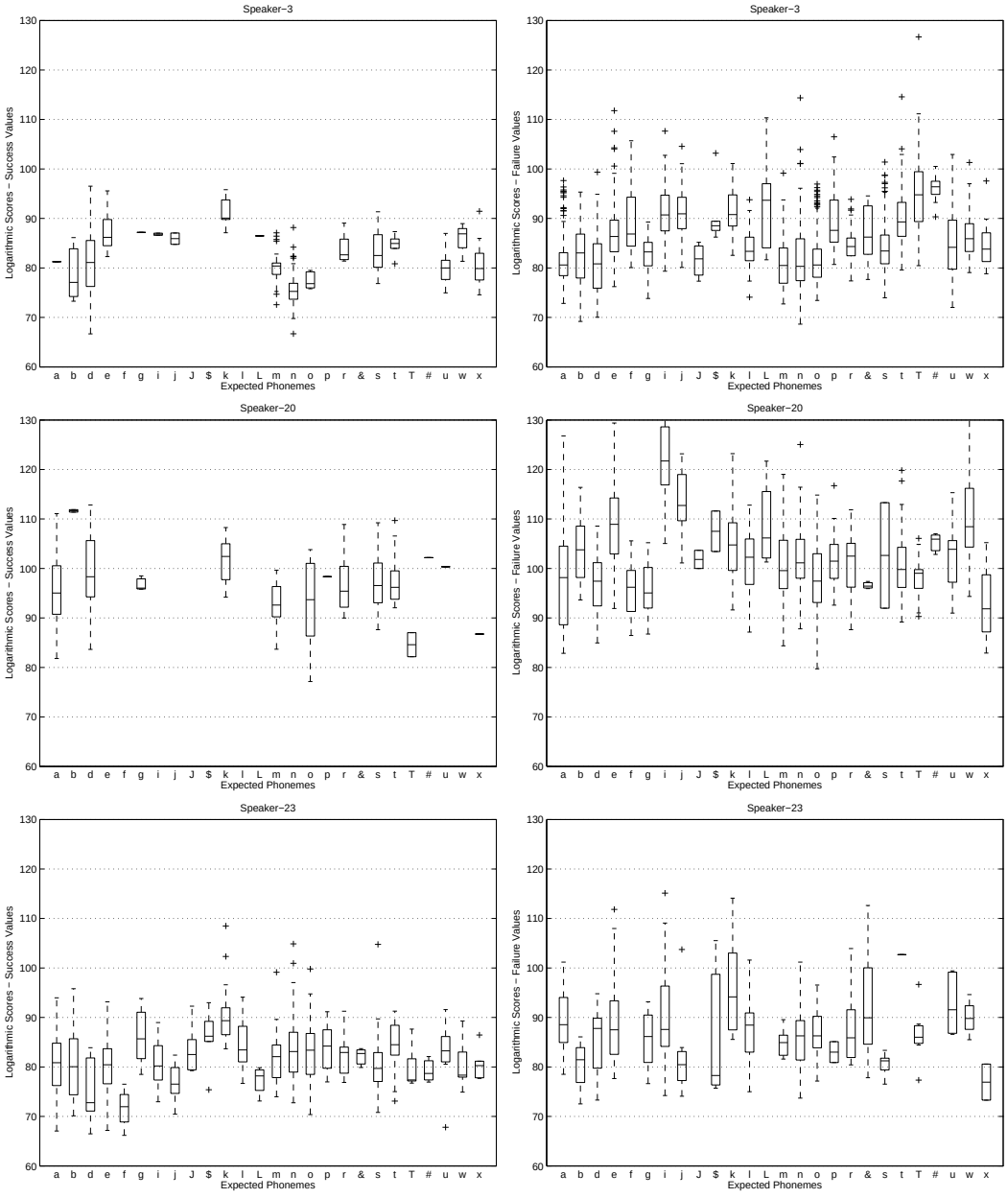


Figure 3.4: **Success and Failure logarithmic scores** for every phoneme. Boxplots comparison of the worst speakers with Albayzin. Row 1: Non-Native m03, row 2: Native L06 and row 3: Albayzin

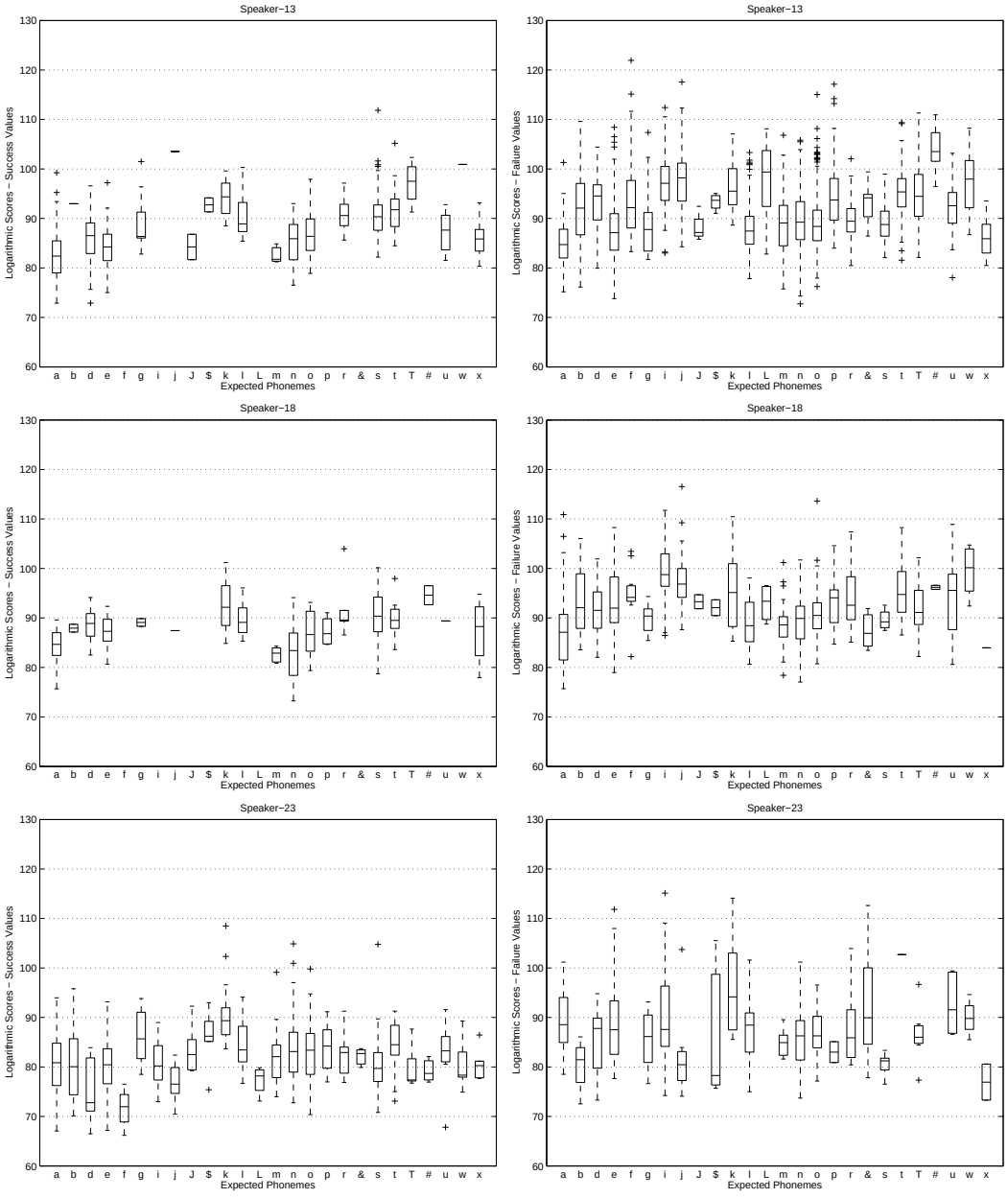


Figure 3.5: **Success and Failure logarithmic scores** for every phoneme. Boxplots comparison of the best speakers with Albayzin. Row 1: Non-Native m03, row 2: Native L06 and row 3: Albayzin

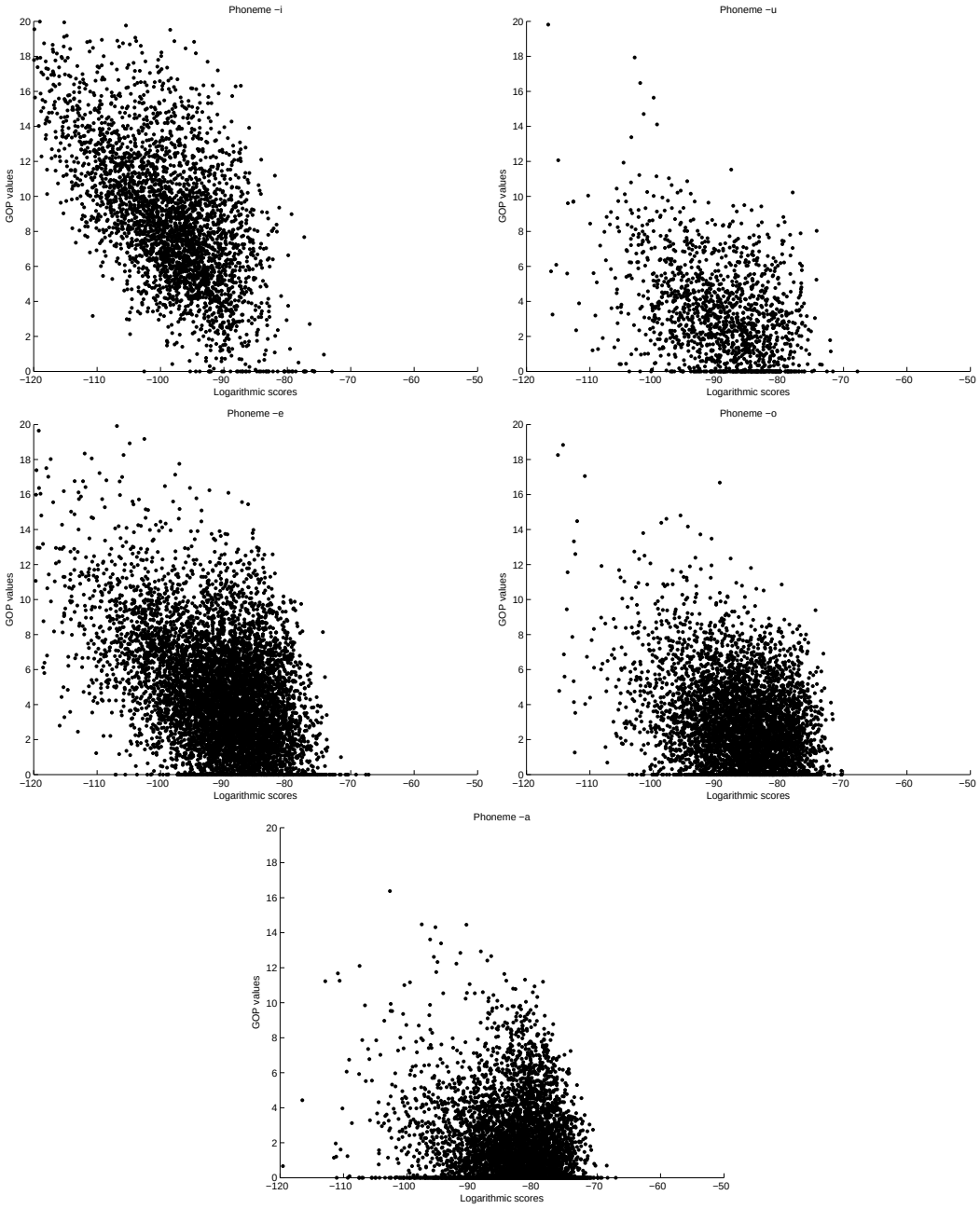


Figure 3.6: Logarithmic scores versus GOP values for vowels. Global Analysis

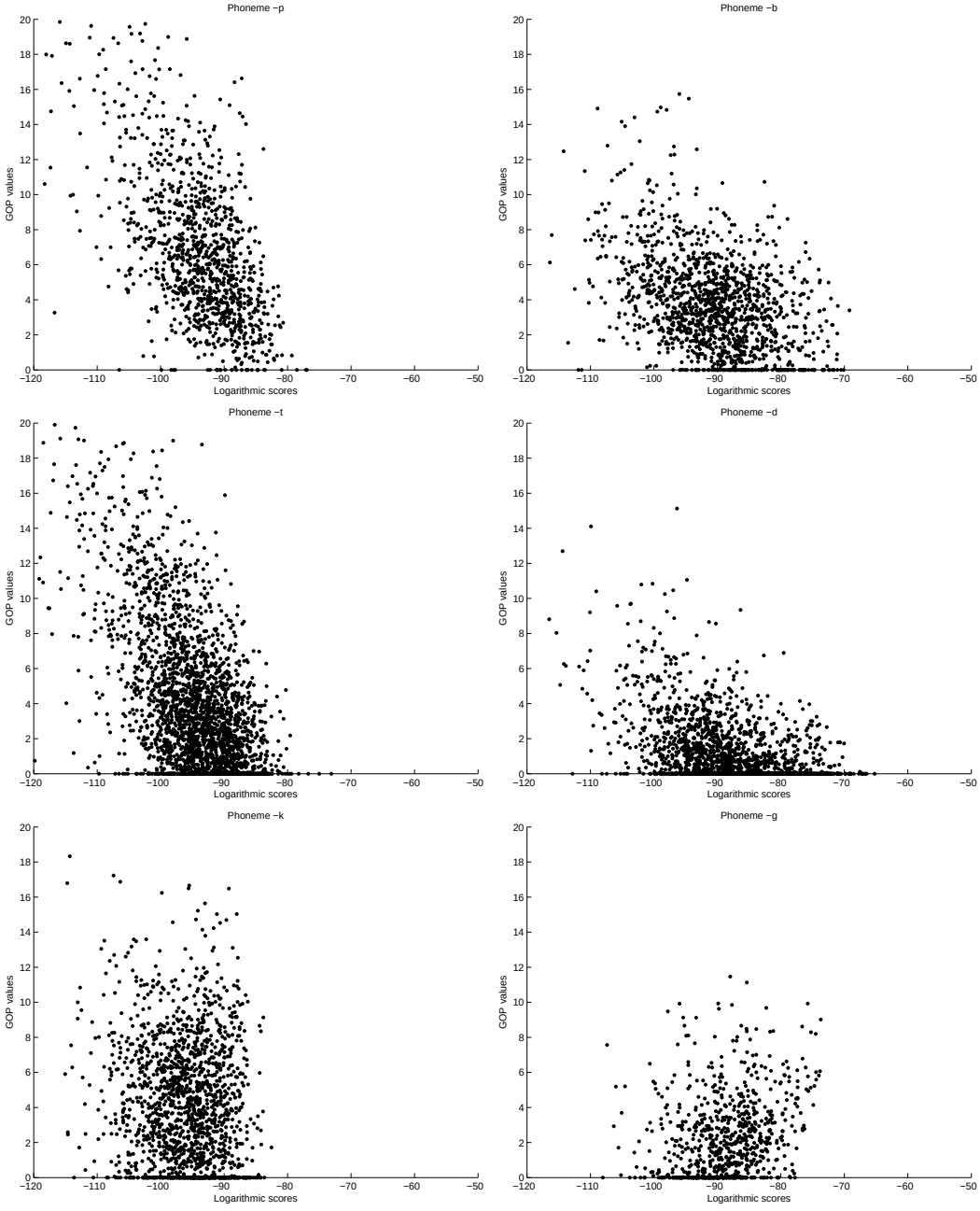


Figure 3.7: Logarithmic scores versus GOP values for plosive consonants . Global Analysis

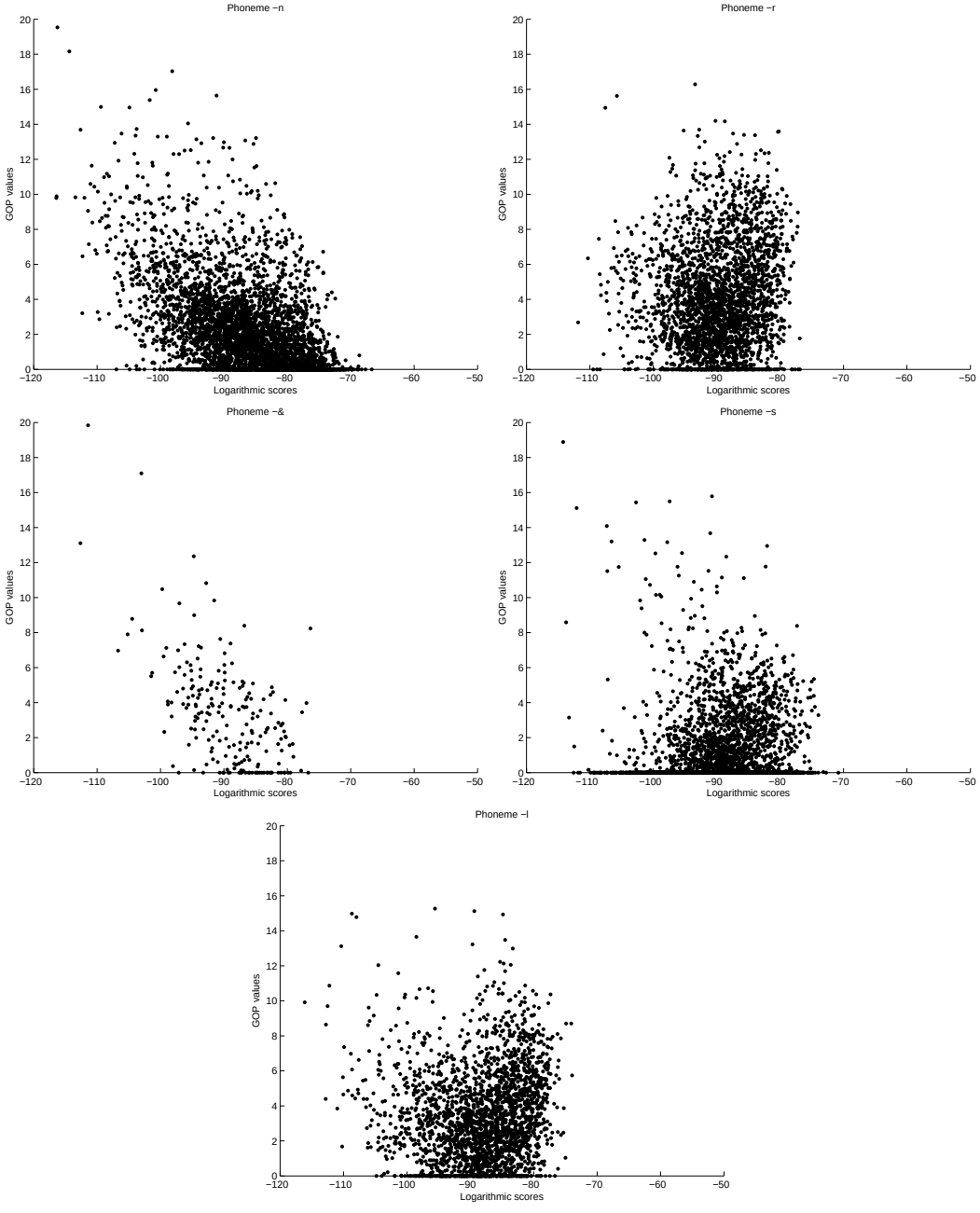


Figure 3.8: Logarithmic scores versus GOP values for alveolar consonants . Global Analysis

3.2.3 *Alternative scoring proposals evaluation*

The results presented in the last sections showed that there is a high dependence on the phoneme. Also, the logarithmic scores depicted the same trend over the whole phonemes with variations on the logarithmic scores.

Based on the latter we decided to create a set of new scorings by computing a new parameter obtained from the logarithmic scores and a set of rules explained next.

Case 1

The computation of a new score n_score for every expected phoneme q_i (that is evaluated in the corpus) is realized by using the logarithmic score ls_{q_i} associated to it and the average y_{m_i} and standard deviation y_{std_i} previously computed.

$$n\_score(q_i) = \frac{|ls_{q_i} - y_{m_i}|}{y_{std_i}} \tag{3.1}$$

The values y_{m_i} and y_{std_i} are obtained by computing all the logarithmic scores of the expected phoneme when uttered by the same speaker and that also comply with some rules for choosing or not the logarithmic score of q_i. These are explained later for every sub-case.

Thus, there is a value of y_{m_i} and y_{std_i} for every phoneme and a given speaker on every sub-case studied, although as we will discuss later depending on the sub-case there are or not y_{m_i} and y_{std_i} values for all phonemes, and therefore some conditions are applied.

- **Case 1.a**

  The case 1.a computes the y_{m_i} and y_{std_i} values over all the ls_{q_i} that correspond to the expected phoneme q_i when its GOP value was equal to zero. As explained before the logarithmic score is only selected if we are analyzing the i-th phoneme.

  For example, by looking table 3.4 suppose we are trying to get the y_{m_i} and y_{std_i} for the expected phoneme $q_i = a$, then the logarithmic score value -82.450127 will be one of the log scores selected, since the GOP value is equal to zero. This selection is evaluated over all utterances of the same speaker to get the logarithmic scores that fit the selection criteria.

  However, as you can see in the boxplot representation of the success logarithmic scores (figures 3.4, 5.9, 5.10, 5.11 and 5.12), for most of the speakers there are no values that can be used for every phoneme, because the speaker didn't uttered correctly the expected phoneme. To overcome this issue we use the Albayzin values, since it is the reference and has y_{m_i} and y_{std_i} values for every phoneme. If only, one

utterance is made for a phoneme then y_{std_i} is changed to one to avoid division over zero.

- **Case 1.b**
  The case 1.b is based on choosing the logarithmic scores that correspond to the expected phoneme q_i when this was uttered by the same speaker, there is no dependency on the GOP value.

- **Case 1.c**
  The case 1.c is based on only choosing the logarithmic scores that correspond to the expected phoneme q_i when its GOP value is different from zero to compute y_{m_i} and y_{std_i}, in opposite to case 1.a.

  Although, for this case there is at least more than one case in which a phoneme was uttered incorrectly then y_{m_i} and y_{std_i} can be obtained. See figures 3.4, 5.9, 5.10, 5.11 and 5.12. In the worst cases if only one utterance is made for a phoneme then y_{std_i} is changed to one to avoid division over zero. Or if no y_{m_i} and y_{std_i} were obtained for a phoneme then the Albayzin values are used.

	ID	CASE 1.a	CASE 1.b	CASE 1.c
Non-Natives	f01	1.92	0.79	0.80
	f02	1.76	0.78	0.80
	m03	2.69	0.77	0.78
	f04	1.53	0.78	0.80
	f05	1.75	0.79	0.81
	f06	1.69	0.79	0.79
	f07	1.55	0.78	0.78
	m08	2.32	0.79	0.80
	f09	1.71	0.77	0.77
	f10	1.77	0.79	0.80
	f11	1.85	0.78	0.75
	f12	2.04	0.78	0.76
	f13	1.76	0.79	0.82
	f14	2.60	0.78	0.77
Natives	L01	2.40	0.78	0.78
	L02	2.58	0.77	0.91
	L03	1.99	0.78	0.82
	L04	2.19	0.79	0.89
	L05	3.43	0.77	0.75
	L06	3.62	0.79	0.75
	L07	2.44	0.79	0.82
	L08	2.23	0.78	0.77
	Albayzin	1.0017	0.7521	1.6104

Table 3.7: Global new score results for every speaker. Cases 1.a, 1.b and 1.c.

Case 2

The case 1 made use of the logarithmic score of the phoneme q_i. The case 2 on the other hand chooses the maximum logarithmic score among all the logarithmic scores calculated for all the phonemes in that specific utterance. With these is possible to calculate the average w_{m_i} and standard deviation w_{std_i} of all the maximum logarithmic scores for every expected phoneme. The w_{m_i} and w_{std_i} values of all phonemes are obtained for every speaker in the whole corpus, meaning that we have an array of 26 rows (phonemes) and 23 columns (speakers) for each of them. This yields the following formula.

For example, by looking table 3.4 if we were trying to obtain the w_{m_i} and w_{std_i} values for the expected phoneme $q_i = a$ then the logarithmic score -82.450127 will be chosen since this is the largest, on the other hand by looking table 3.5 if we were trying to obtain the w_{m_i} and w_{std_i} values for the expected phoneme $q_i = n$ then the logarithmic score -85.112442 will be chosen since this is the largest even when it doesn't correspond to the expected phoneme.

$$n\_score(q_i) = \frac{|ls_{q_i} - w_{m_i}|}{w_{std_i}} \tag{3.2}$$

The case 2 is sub-divided into two, whose difference is basically using or removing the absolute value from the numerator in equation 3.2.

- **Case 2.a**

$$n\_score(q_i) = \frac{|ls_{q_i} - w_{m_i}|}{w_{std_i}} \tag{3.3}$$

- **Case 2.b**

$$n\_score(q_i) = \frac{ls_{q_i} - w_{m_i}}{w_{std_i}} \tag{3.4}$$

Once all these new scores have been obtained for every expected phoneme in the corpus, a global new pronunciation score is obtained for every speaker.

$$gnew_s = \frac{\sum_{n=1}^{N} n\_score(q_{i_n})}{N} \tag{3.5}$$

Where N is the total of all phonemes uttered by the given speaker. See tables 3.7 and 3.8 to check the results.

	ID	CASE 2.a	CASE 2.b
Non-Natives	f01	1.04	-0.62
	f02	1.06	-0.68
	m03	1.31	-1.15
	f04	1.13	-0.81
	f05	1.07	-0.74
	f06	1.04	-0.70
	f07	1.09	-0.80
	m08	1.08	-0.76
	f09	1.05	-0.69
	f10	1.13	-0.79
	f11	1.11	-0.70
	f12	1.01	-0.58
	f13	0.98	-0.58
	f14	1.06	-0.65
Natives	L01	1.19	-0.76
	L02	1.09	-0.69
	L03	1.11	-0.69
	L04	1.03	-0.58
	L05	1.20	-0.84
	L06	1.25	-0.82
	L07	1.23	-0.81
	L08	1.20	-0.82
	Albayzin	0.8368	-0.1449

Table 3.8: Global new score results for every speaker. Cases 2.a, 2.b and 2.c

Scorings Comparison

The results obtained from the different automated global scorings detailed before are now contrasted among them and with the human judgments. We have chosen two of the perceptual dimensions scorings used in [1]. First, the phonetic correctness (HJ PHO), on which the experts evaluated in a scale from 1 to 5 (1:clearly non-native and 5:native) how well pronounced were the phonemes. Second, the Spanish level (DELE), on which the experts indicated the level of Spanish proficiency (A1, A2, ... C2) based on a numeric scale from 1 to 6 respectively.

The table 3.9 presents all the scorings. The human judgments scores for every speaker are the average of all the scores given by the experts, these are presented in the second and third column. The higher is the score means that the speaker has a higher level in that perceptual dimension. On the other hand, the automated scores tell that a speaker has better quality of pronunciation if its score is closer to zero, except for case 1.c which is inverse, the score is better the farther from zero.

The pairwise Pearson correlation between all the different scorings was evaluated at speaker level. Results for the non-natives are presented in table 3.10 and for all the speakers (without Albayzin) the results are presented in table 3.11.

	ID	HJ PHO	HJ DELE	GOP	CASE 1.a	CASE 1.b	CASE 1.c	CASE 2.a	CASE 2.b
Non-Natives	f01	3.47	3.91	3.52	1.92	0.79	0.80	1.04	-0.62
	f02	3.22	3.53	3.75	1.76	0.78	0.80	1.06	-0.68
	m03	3.08	3.09	5.14	2.69	0.77	0.78	1.31	-1.15
	f04	3.13	3.32	3.77	1.53	0.78	0.80	1.13	-0.81
	f05	3.42	3.68	3.55	1.75	0.79	0.81	1.07	-0.74
	f06	3.19	3.41	3.75	1.69	0.79	0.79	1.04	-0.70
	f07	3.30	3.50	4.03	1.55	0.78	0.78	1.09	-0.80
	m08	3.23	3.23	3.95	2.32	0.79	0.80	1.08	-0.76
	f09	3.10	3.43	3.98	1.71	0.77	0.77	1.05	-0.69
	f10	3.23	3.67	4.35	1.77	0.79	0.80	1.13	-0.79
	f11	2.92	2.86	3.81	1.85	0.78	0.75	1.11	-0.70
	f12	3.01	3.28	3.00	2.04	0.78	0.76	1.01	-0.58
	f13	3.77	3.92	2.96	1.76	0.79	0.82	0.98	-0.58
	f14	3.12	3.18	3.17	2.60	0.78	0.77	1.06	-0.65
Natives	L01	4.95	5.00	3.57	2.40	0.78	0.78	1.19	-0.76
	L02	4.95	5.00	3.10	2.58	0.77	0.91	1.09	-0.69
	L03	5.00	5.00	3.64	1.99	0.78	0.82	1.11	-0.69
	L04	5.00	5.00	2.72	2.19	0.79	0.89	1.03	-0.58
	L05	5.00	5.00	4.46	3.43	0.77	0.75	1.20	-0.84
	L06	4.85	5.00	4.46	3.62	0.79	0.75	1.25	-0.82
	L07	4.95	5.00	3.84	2.44	0.79	0.82	1.23	-0.81
	L08	5.00	5.00	4.11	2.23	0.78	0.77	1.20	-0.82
	Albayzin	-	-	0.8702	1.0017	0.7521	1.6104	0.8368	-0.1449

Table 3.9: Comparison of all the different scores analyzed

Results for non-natives show that the automated scoring proposal of case 1.c correlates well with the human judgments, although it has almost no correlation with the GOP metric. The GOP metric has little correlation with the human judgments but is highly correlated with cases 2.a and 2.b probably because both depend on the use of the maximum logarithmic scores.

However, results differ for correlations among scorings when all the speakers are taken into account. The case 1.a shows the best correlation values with the human judgments. The GOP metric shows practically no correlation with the human judgments, but it keeps maintaining a high correlation with cases 2.a and 2.b. The case 1.c has lower correlation than when is only computed for the non-natives. In general, all the automated scorings show low correlation with the human judgments but this is attributable to the almost non variation among the human judgments values for the native speakers, see columns 2 and 3 in table 3.9.

Non-Native Speakers								
	HJ PHO	**HJ DELE**	**GOP**	**Case 1.a**	**Case 1.b**	**Case 1.c**	**Case 2.a**	**Case 2.b**
HJ PHO	1							
HJ DELE	0.87	1						
GOP	-0.35	-0.31	1					
Case 1.a	-0.26	-0.45	0.22	1				
Case 1.b	0.50	0.49	-0.34	-0.13	1			
Case 1.c	**0.81**	**0.74**	-0.03	-0.23	0.49	1		
Case 2.a	-0.41	-0.46	**0.88**	0.42	-0.38	-0.09	1	
Case 2.b	0.30	0.38	**-0.90**	-0.36	0.37	-0.01	-0.96	1

Table 3.10: **Correlation coefficients** among pronunciation scores at speaker level for **non-native speakers**.

All Speakers								
	HJ PHO	**HJ DELE**	**GOP**	**Case 1.a**	**Case 1.b**	**Case 1.c**	**Case 2.a**	**Case 2.b**
HJ PHO	1							
HJ DELE	0.99	1						
GOP	*-0.09*	-0.1	1					
Case 1.a	**0.55**	0.51	0.32	1				
Case 1.b	-0.07	-0.03	-0.11	-0.18	1			
Case 1.c	0.37	0.38	-0.51	-0.14	-0.08	1		
Case 2.a	0.39	0.35	**0.76**	0.62	-0.15	-0.26	1	
Case 2.b	-0.03	0.02	**-0.88**	-0.37	0.22	0.28	-0.87	1

Table 3.11: **Correlation coefficients** among pronunciation scores at speaker level for **all speakers**.

Chapter 4

Conclusions and Future Work

4.1 *Conclusions*

The previous work consisted of the analysis over the GOP scoring results on a subset of the SAMPLE corpus to detect its weaknesses and strengths, and therefore suggest improvements and also create new scoring metrics. As described before, the results were evaluated from different perspectives, at phoneme level and speaker level over the GOP values and the logarithmic scores.

Through the GOP results analysis we noted that this metric generated results that were not as consistent as desired when grading all the speakers, although the worst and best speakers among non-natives were identified correctly. When looking at the GOP values distributions at phoneme level the following conclusions arose:

- The best speakers have phoneme-GOP distributions closer to zero and with less deviation than the worst speakers.
- The best speakers (whether natives or not) also demonstrate better results for certain phonemes (distributions always closer to zero), while the worst speakers show difficulties with them (phoneme *a* for example). This means that there is a dependency at phoneme level that can describe the pronunciation level of the speaker.
- Some phonemes are more problematic for the non-natives than for natives (phonemes T and f) have worst distributions for non-natives (even for the best speaker) than for native speakers.

However, the results were not good for natives when compared to the non-natives results, we would have expected that the GOP global score values for natives were lower than the non-native scores. This lead us to the analysis of the logarithmic scores to try understanding if this issue could be related or not to its computation.

The logarithmic scores distributions of all the speakers marked a similar trend depicting a phoneme dependency. The distributions showed that the natives had logarithmic scores that were farther from zero (more negative) than the non-natives, which takes us to the following conclusion:

- The log-likelihoods obtained for the natives are farther from zero than the reference and non-natives, meaning that the acoustic segments of the natives doesn't fit well with the models, therefore a re-evaluation of these is suggested.

The evaluation of the success and failure cases over the logarithmic scores showed that the number of correct uttered phonemes is directly related to the pronunciation level of the speaker. The speakers with poorest pronunciation quality are the ones that uttered less correct phonemes and whose distributions have greater variations.

Likewise, certain groups of phonemes where evaluated by checking the relation of their logarithmic scores with their GOP score, these results confirmed also that certain phonemes are more problematic than others.

Finally, we saw that computing new metrics that were based on the logarithmic scores could probably yield good results, because according to our previous analysis these depicted a lot of information that could also tell the speakers level.

Based on the five different cases and its comparison with the human judgments, the next conclusions are given:

- Using the logarithmic scores to obtain a new normalized parameter at speaker level yields a scoring metric whose performance is comparable or even better than the GOP.

- There is a strong dependency among the average logarithmic scores and the speaker. This represents a problem to try establishing adequate and generalized thresholds that are suitable for a correct evaluation of any given speaker.

- The case 1.c showed a high correlation with the human judgments, this probably happens because the new computed parameter (failure cases) maximizes the correct utterances and diminish the incorrect.

- Using the maximum logarithmic scores to compute the new parameter creates new metric whose results highly correlate with the GOP, meaning there is no great variation when choosing an instantaneous max log score or the new parameter.

Unfortunately the automated metrics cannot be correctly compared for the native speakers because the human judgments are not good referents of their pronunciation quality. Also, for all the automated metrics, the scoring values of the native speakers were worst than the non-native which is probably caused by errors in the data used to compute the logarithmic scores, further research should be made to detect the real causes.

4.2 *Future Work*

The results obtained in this work evidence the necessity of continuing with this line of research, some future work suggestions are stated next:

- Extend the SAMPLE corpus to have more native and non-native speakers that will allow to better evaluate correlation among human judgments and the scoring metrics evaluated.
- Redefine the way in which the experts grade the quality of pronunciation at phonetic level to have more accurate and discriminant scores.
- Review the data used to compute the logarithmic scores for the native speakers due to the problems encountered before.
- Compute the new metrics to obtain a score at phrase level that can be contrasted with the human judgments at phrase level that are already available.
- Use the new metrics to identify phonemes whose pronunciation is more difficult according to the type of speaker (non-native language).

Bibliography

[1] V. Cardeñoso-Payo, C. González-Ferreras, and D. Escudero-Mancebo. Assessment of non-native prosody for spanish as l2 using quantitative scores and perceptual evaluation.

[2] Mario Carranza. Errores y dificultades específicas en la adquisición de la pronunciación del español le por hablantes de japonés y propuestas de corrección. *Nuevos enfoques en la enseñanza del español en Japón–Concha Moreno y GIDE–.Tokyo: Asahi*, 2012.

[3] Jonathan D. Culler. *Ferdinand de Saussure*. Cornell University Press, 1986.

[4] Maxine Eskenazi. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844, 2009.

[5] Horacio Franco, Leonardo Neumeyer, Yoon Kim, and Orith Ronen. Automatic pronunciation scoring for language instruction. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1471–1474. IEEE, 1997.

[6] Xavier Frías Conde. Introducción a la fonética y fonología del español. *el suplemento*, 2001.

[7] Juan M. Garrido, David Escudero, Lourdes Aguilar, Valentín Cardeñoso, Emma Rodero, Carme De-La-Mota, César González, Carlos Vivaracho, Sílvia Rustullet, and Olatz Larrea. Glissando: a corpus for multidisciplinary prosodic studies in spanish and catalan. *Language resources and evaluation*, 47(4):945–971, 2013.

[8] Sandra Kanters, Catia Cucchiarini, and Helmer Strik. The goodness of pronunciation algorithm: a detailed performance study. *SLaTE*, 2009:2–5, 2009.

[9] Yoon Kim, Horacio Franco, and Leonardo Neumeyer. Automatic pronunciation scoring of specific phone segments for language instruction. In *Eurospeech*, 1997.

[10] Dean Luo, Yu Qiao, Nobuaki Minematsu, Yutaka Yamauchi, and Keikichi Hirose. Analysis and utilization of mllr speaker adaptation technique for learners' pronunciation evaluation. In *INTERSPEECH*, pages 608–611, 2009.

[11] Brian Mak, Manhung Siu, Mimi Ng, Yik-Cheung Tam, Yu-Chung Chan, Kin-Wah Chan, Ka-Yee Leung, Simon Ho, Fong-Ho Chong, and Jimmy Wong. Plaser: pronunciation learning via automatic speech recognition. In *Proceedings of the*

HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2, pages 23–29. Association for Computational Linguistics, 2003.

[12] Asunción Moreno, Dolors Poch, Antonio Bonafonte, Eduardo Lleida, Joaquim Llisterri, Jose B. Marino, and Climent Nadeu. Albayzin speech database: design of the phonetic corpus. In *Third European Conference on Speech Communication and Technology, EUROSPEECH 1993, Berlin, Germany, September 22-25, 1993*, 1993.

[13] Ambra Neri, Catia Cucchiarini, and H. Strik. Automatic speech recognition for second language learning: How and why it actually works. In *Proc. ICPhS*, pages 1157–1160, 2003.

[14] Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub. Automatic scoring of pronunciation quality. *Speech Communication*, 30(2):83–93, 2000.

[15] Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price. Automatic text-independent pronunciation scoring of foreign language student speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1457–1460. IEEE, 1996.

[16] Antonio Quilis. *Principios de fonología y fonética españolas*, volume 43. Arco libros, 1997.

[17] Helmer Strik, Khiet Truong, Febe De Wet, and Catia Cucchiarini. Comparing different approaches for automatic pronunciation error detection. *Speech Communication*, 51(10):845–852, 2009.

[18] Tomás N. Tomás. *Manual de pronunciación española*. Number 3. Editorial CSIC-CSIC Press, 1991.

[19] Joost van Doremalen, Catia Cucchiarini, and Helmer Strik. Automatic pronunciation error detection in non-native speech: The case of vowel errors in dutch. *The Journal of the Acoustical Society of America*, 134(2):1336–1347, 2013.

[20] Gethin Williams and Steve Renals. Confidence measures for hybrid hmm/ann speech recognition. 1997.

[21] Silke Witt and Steve Young. Computer-assisted pronunciation teaching based on automatic speech recognition. *Language Teaching and Language Technology Groningen, The Netherlands*, 1997.

[22] Silke M. Witt. Automatic error detection in pronunciation training: Where we are and where we need to go. *Proc.IS ADEPT*, 2012.

[23] Silke M. Witt and Steve J. Young. Language learning based on non-native speech recognition. In *Eurospeech*. Citeseer, 1997.

[24] Silke M. Witt and Steve J. Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2):95–108, 2000.

[25] Silke Maren Witt. *Use of speech recognition in computer-assisted language learning*. PhD thesis, University of Cambridge, 1999.

Chapter 5

Appendix A

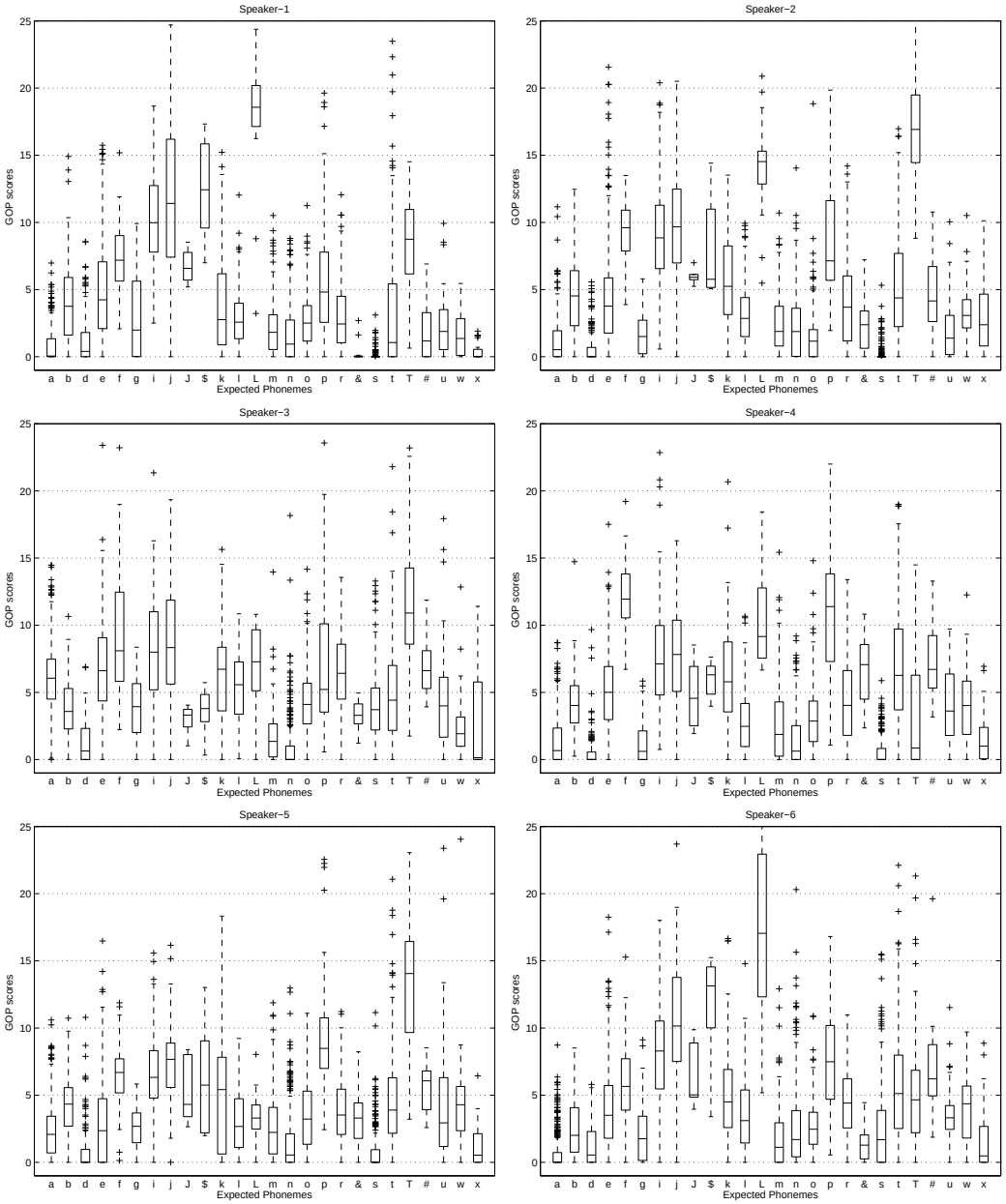


Figure 5.1: Boxplot representation of the statistical data of the **GOP scores** for every expected phoneme. Non-Native Speakers (f01, f02, m03, f04, f05, f06)

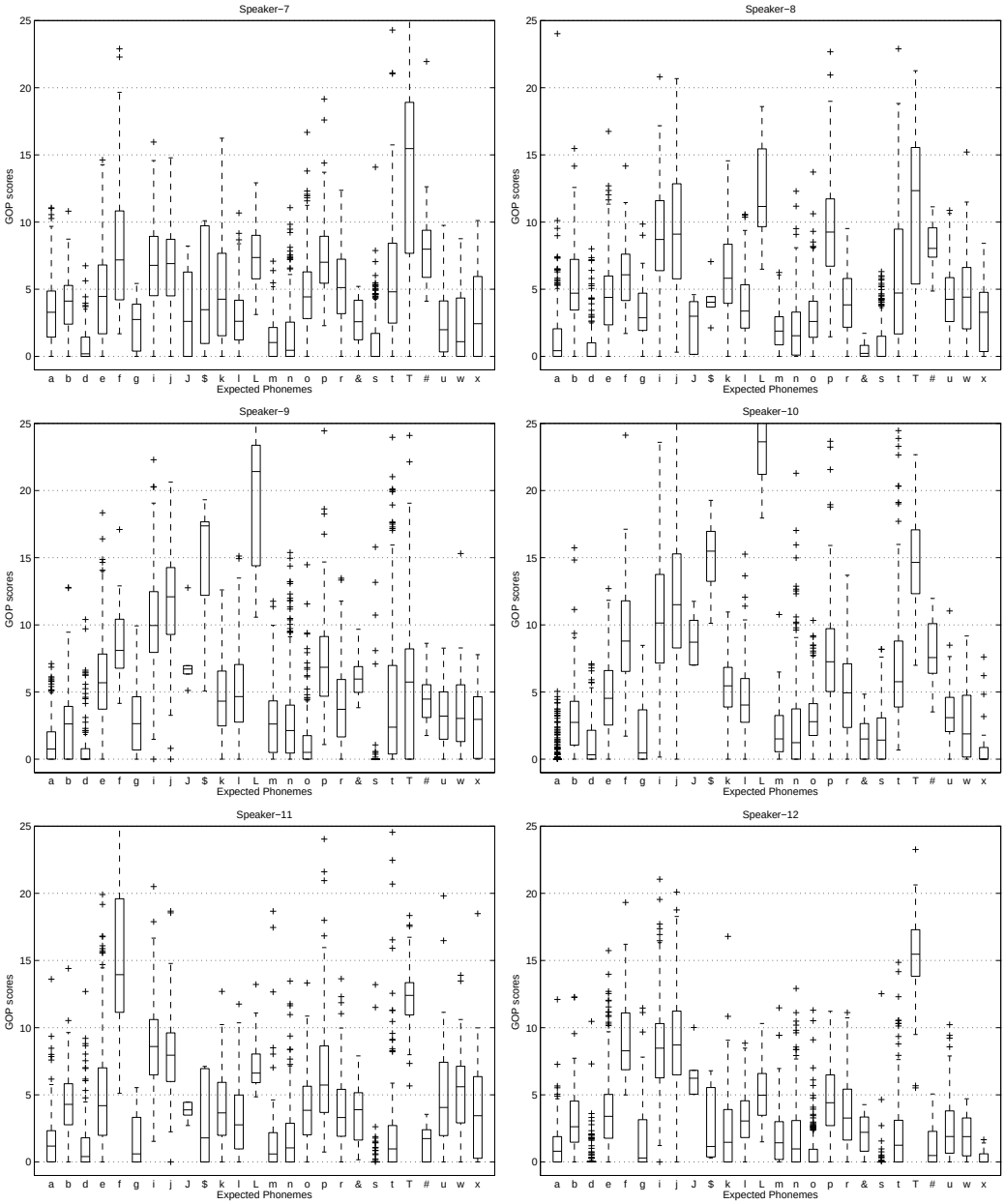


Figure 5.2: Boxplot representation of the statistical data of the **GOP scores** for every expected phoneme. Non-Native Speakers (f07, m08, f09, f10, f11, f12).

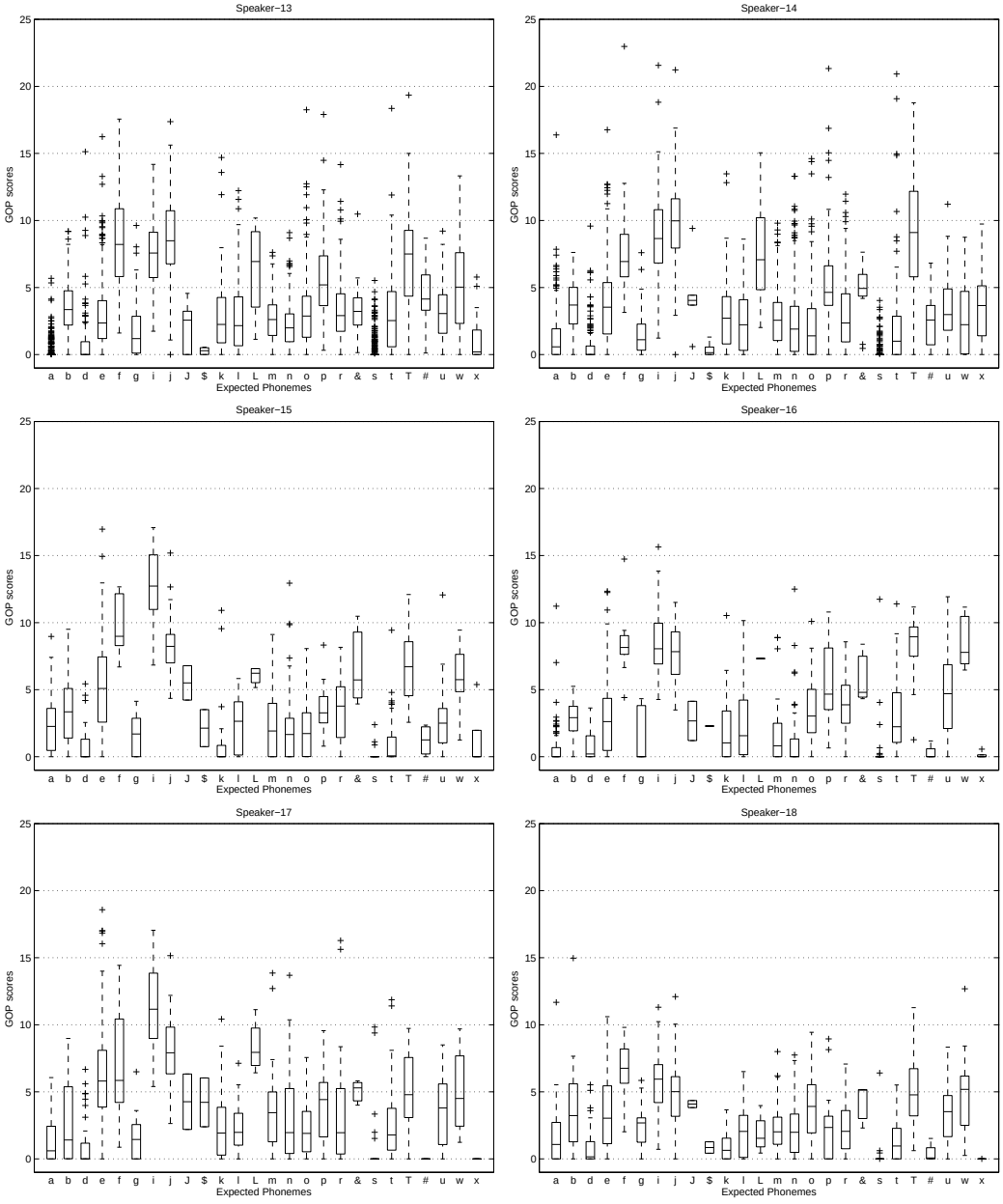


Figure 5.3: Boxplot representation of the statistical data of the **GOP scores** for every expected phoneme. Non-Native Speakers(f13, f14) and Native Speakers (L01, L02, L03, L04).

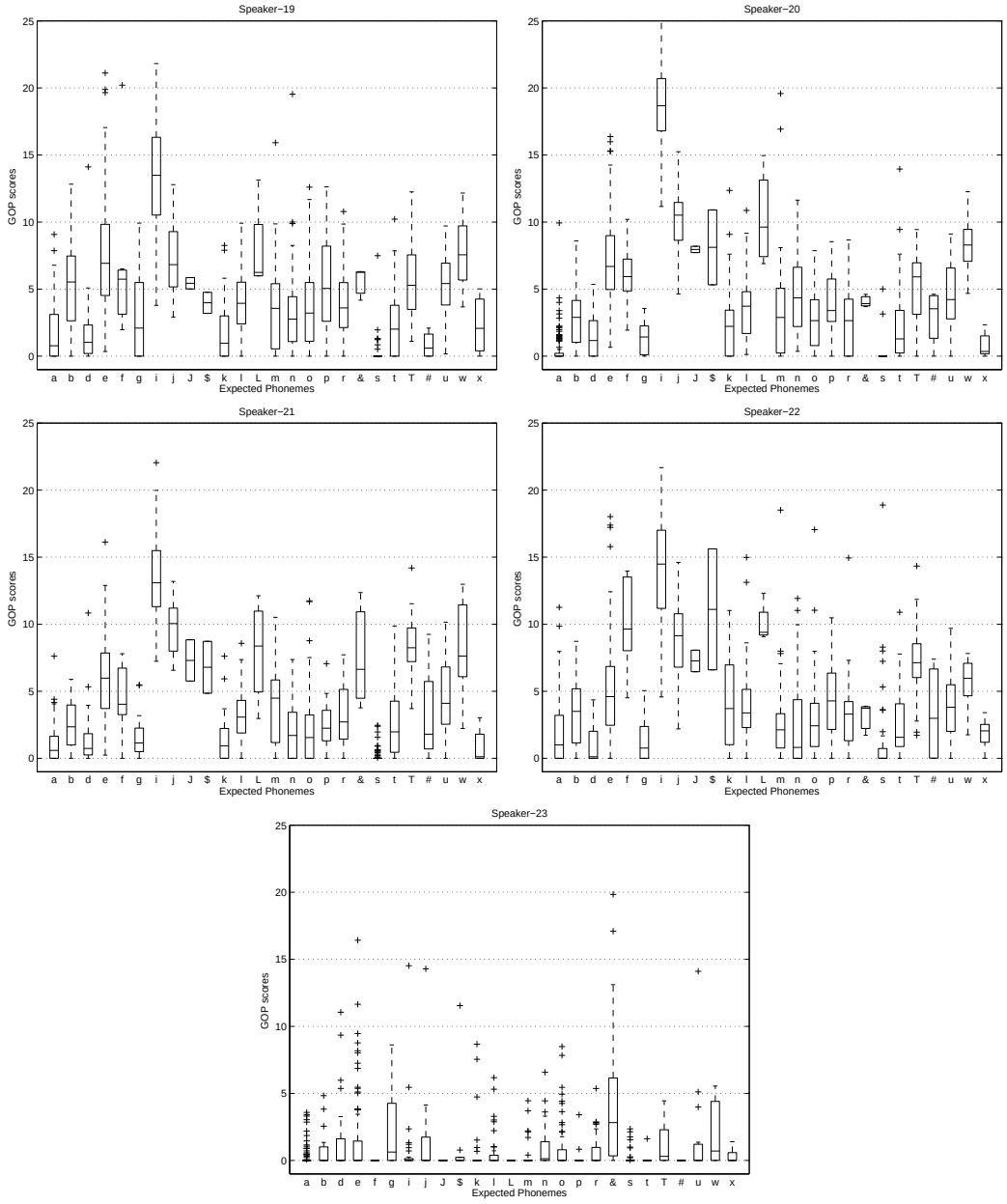


Figure 5.4: Boxplot representation of the statistical data of the **GOP scores** for every expected phoneme. Native Speakers (L05, L06, L07, L08) and Albayzin.

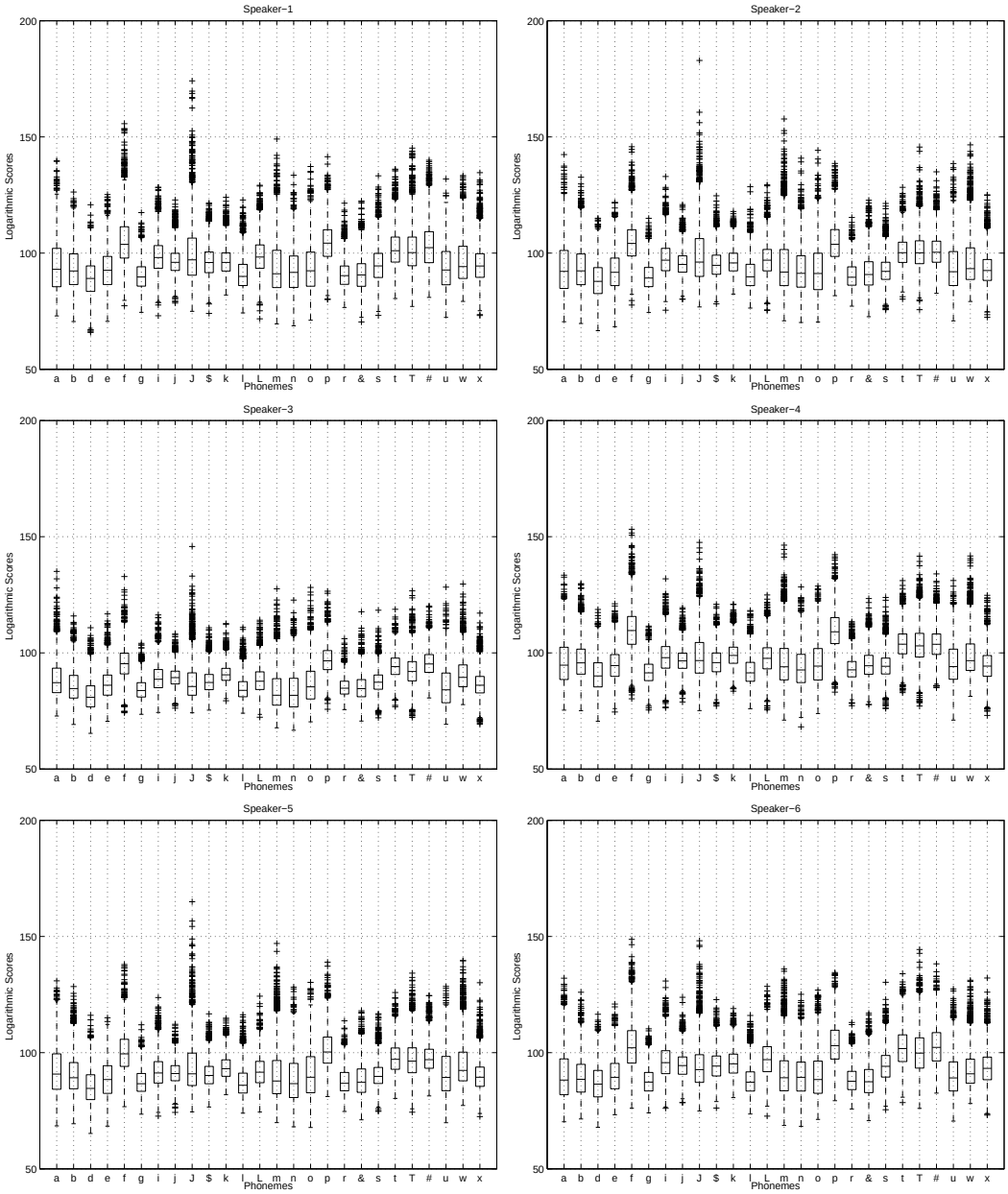


Figure 5.5: Boxplot representation of the statistical data of the **logarithmic scores** for every phoneme. Non-Natives (f01, f02, m03, f04, f05, f06).

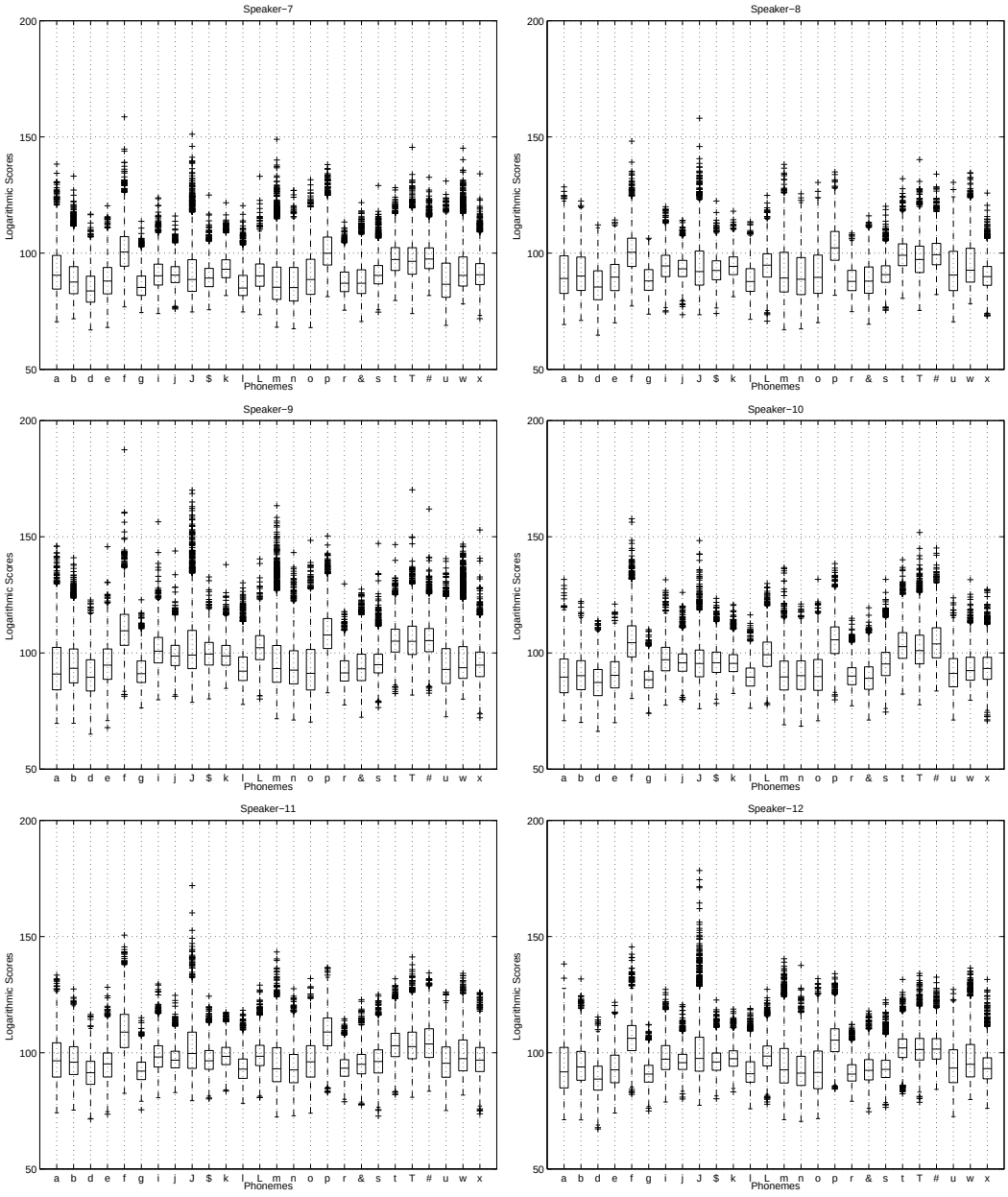


Figure 5.6: Boxplot representation of the statistical data of the **logarithmic scores** for every phoneme. Non-Natives (f07, m08, f09, f10, f11, f12).

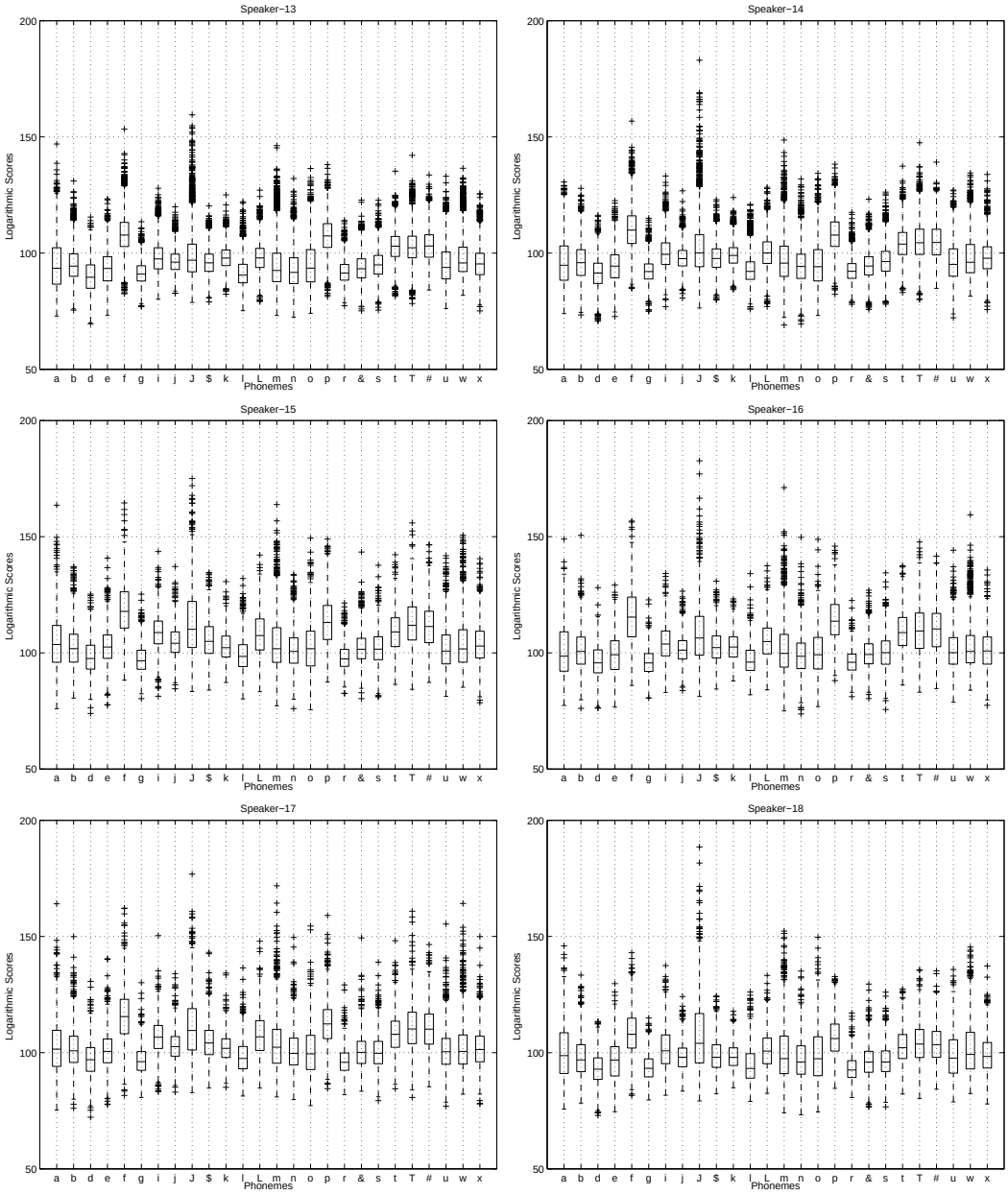


Figure 5.7: Boxplot representation of the statistical data of the **logarithmic scores** for every phoneme. Non-Natives (f13, f14) and Natives (L01, L02, L03, L04).

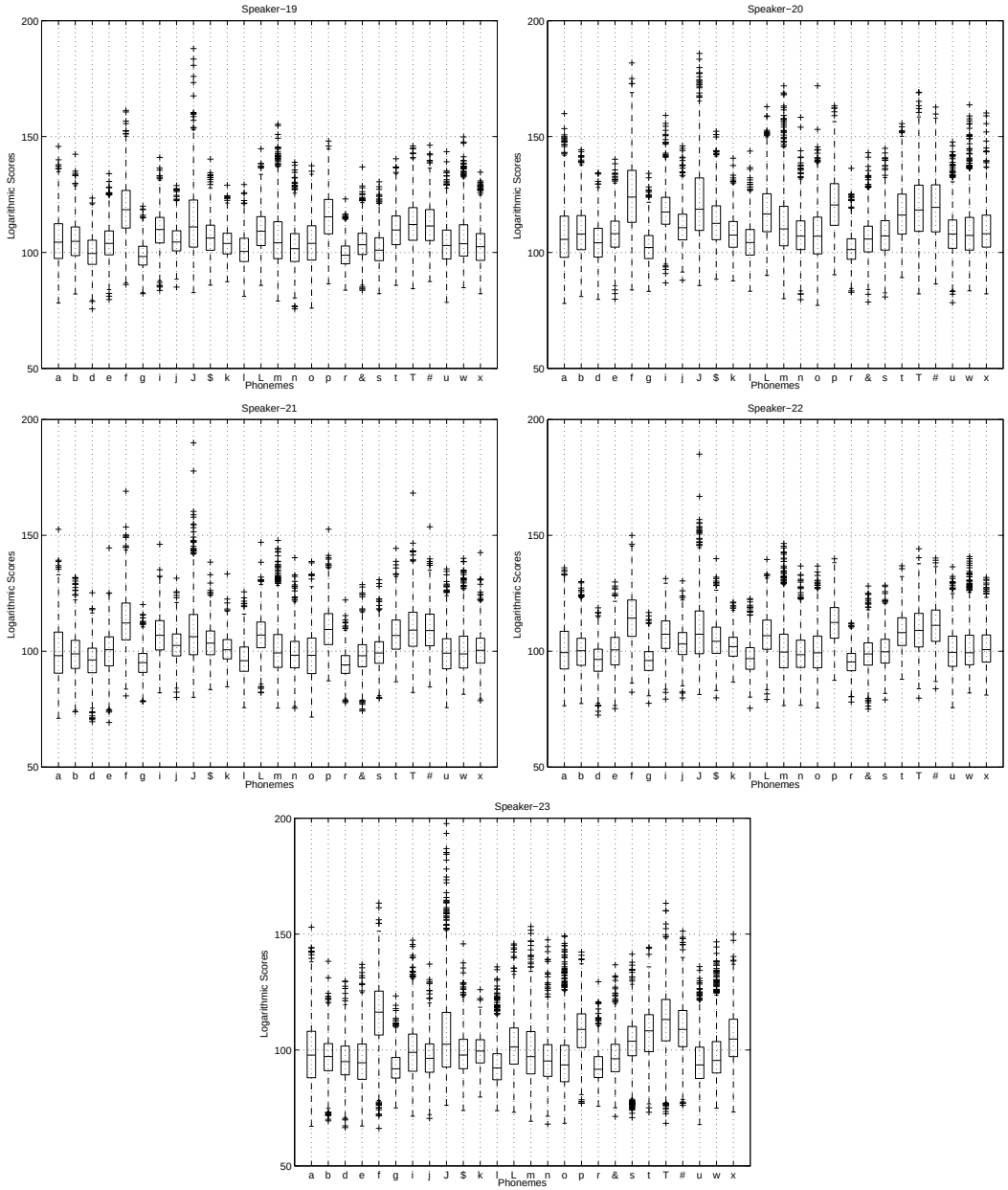


Figure 5.8: Boxplot representation of the statistical data of the **logarithmic scores** for every phoneme. Natives (L05, L06, L07, L08) and Albayzin.

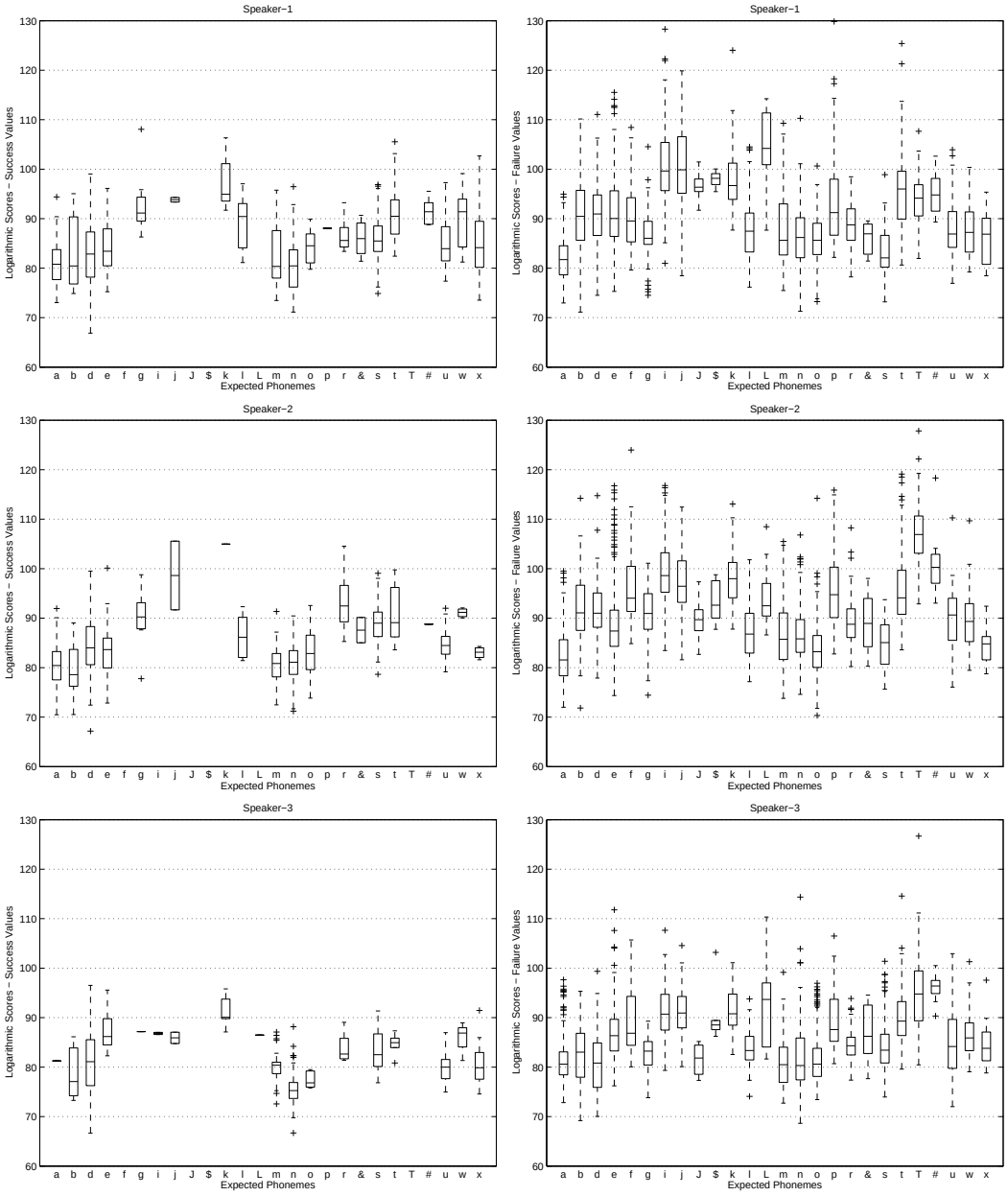


Figure 5.9: **Success and Failure logarithmic scores** for every phoneme. Boxplots comparison. Speakers f01,f02,m03.

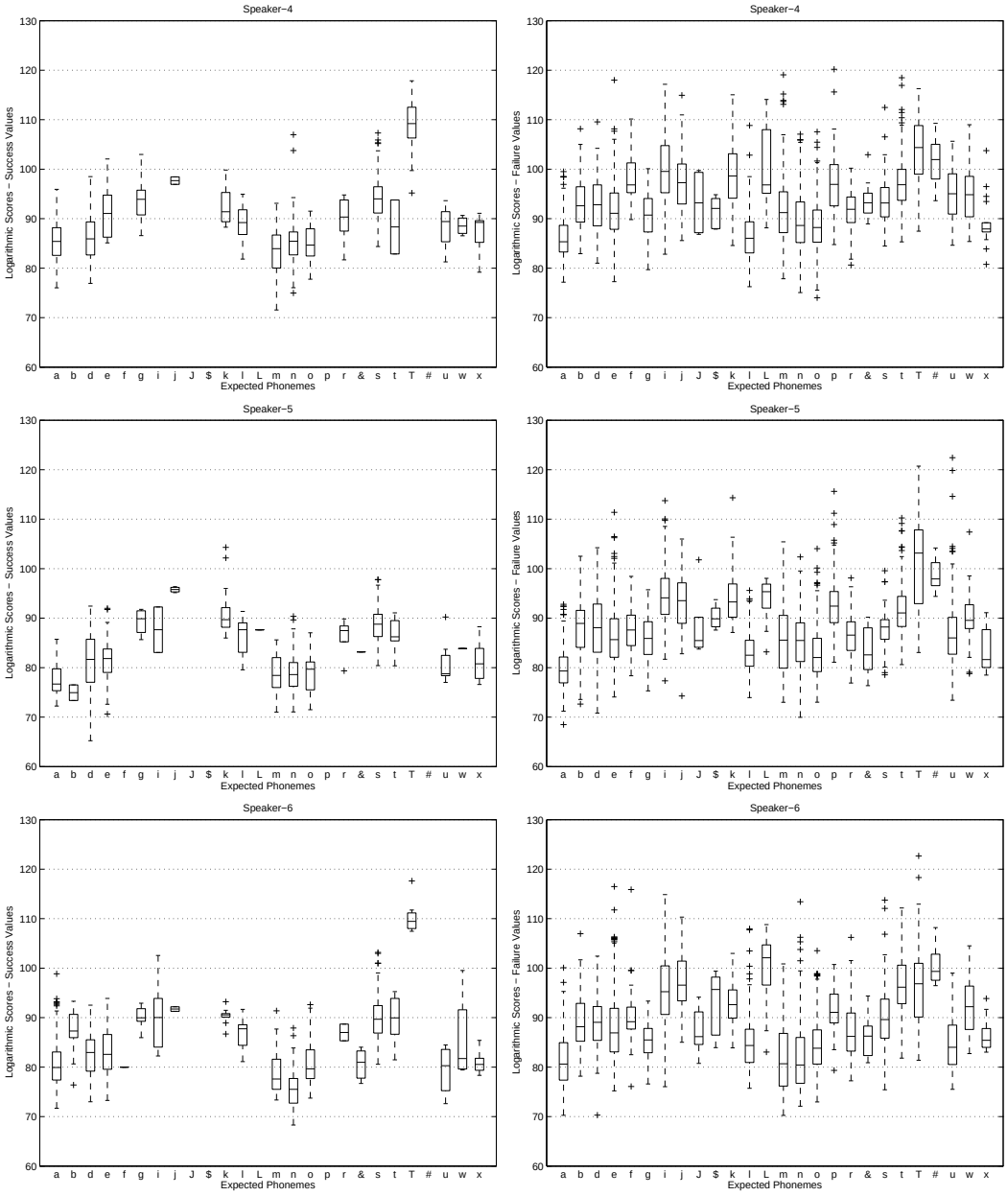


Figure 5.10: **Success and Failure logarithmic scores** for every phoneme. Boxplots comparison. Speakers f04,f05,f06.

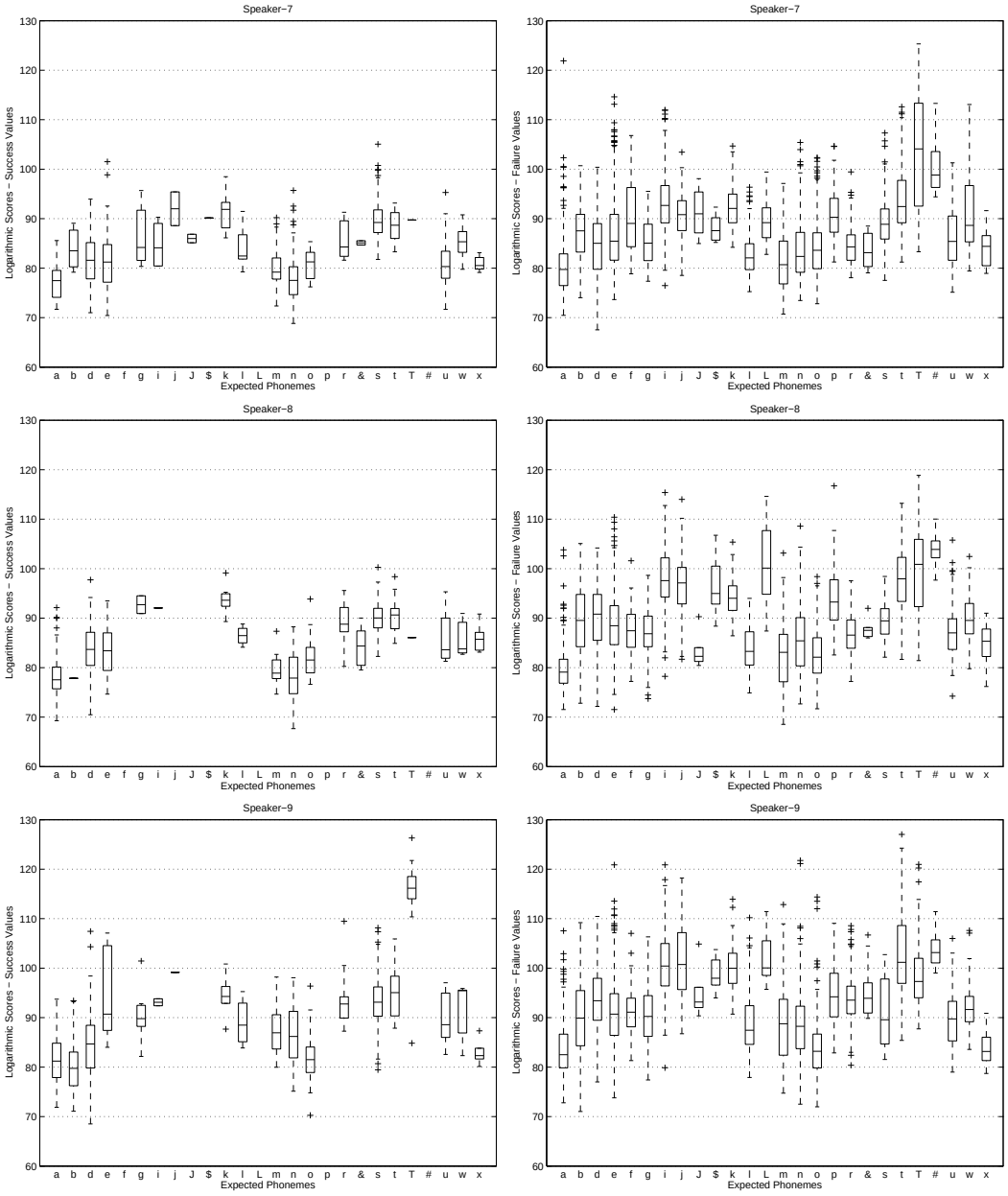


Figure 5.11: **Success and Failure logarithmic scores** for every phoneme. Boxplots comparison. Speakers f07, m08, f09.

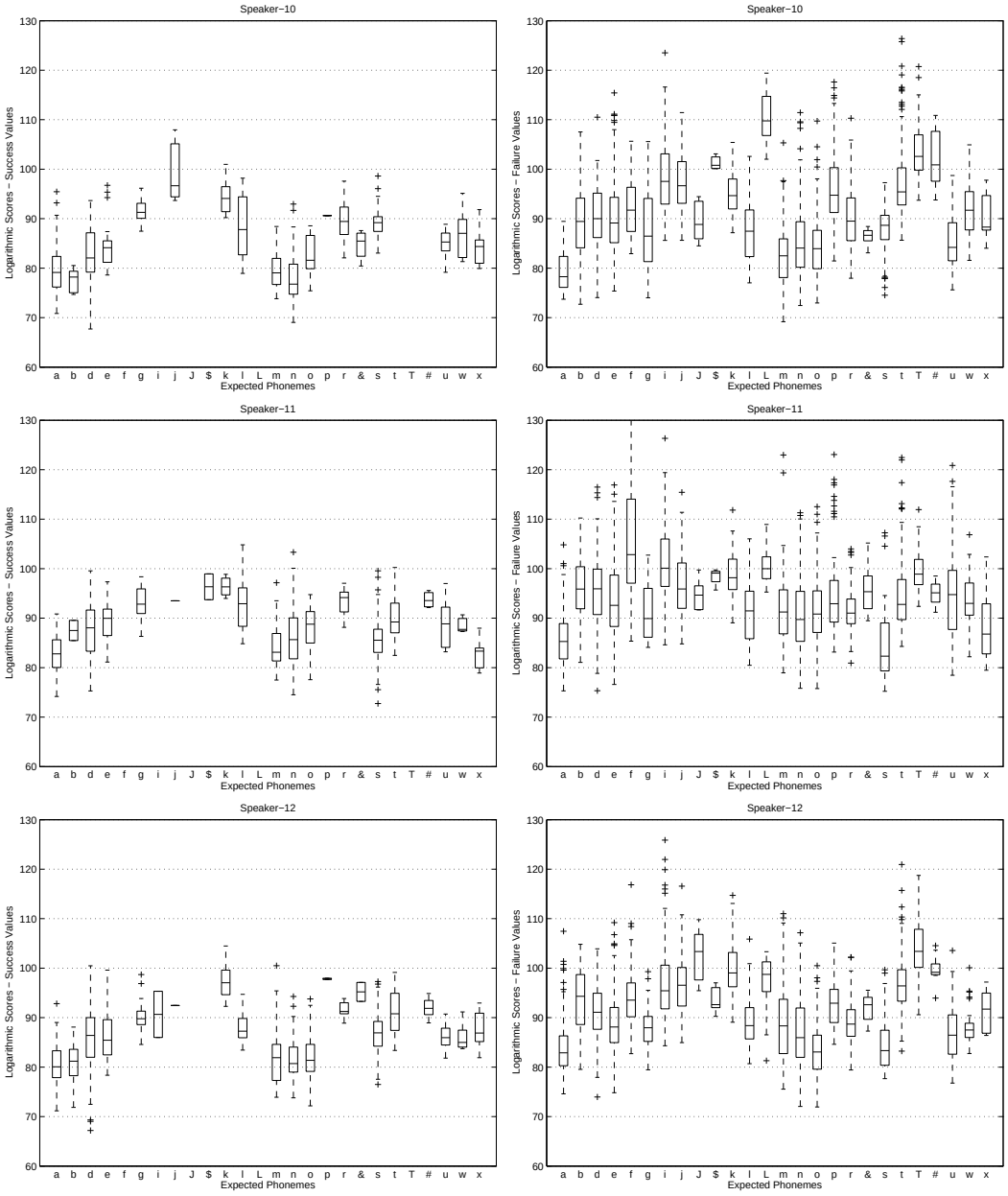


Figure 5.12: **Success and Failure logarithmic scores** for every phoneme. Boxplots comparison. Speakers f10, f11, f12.

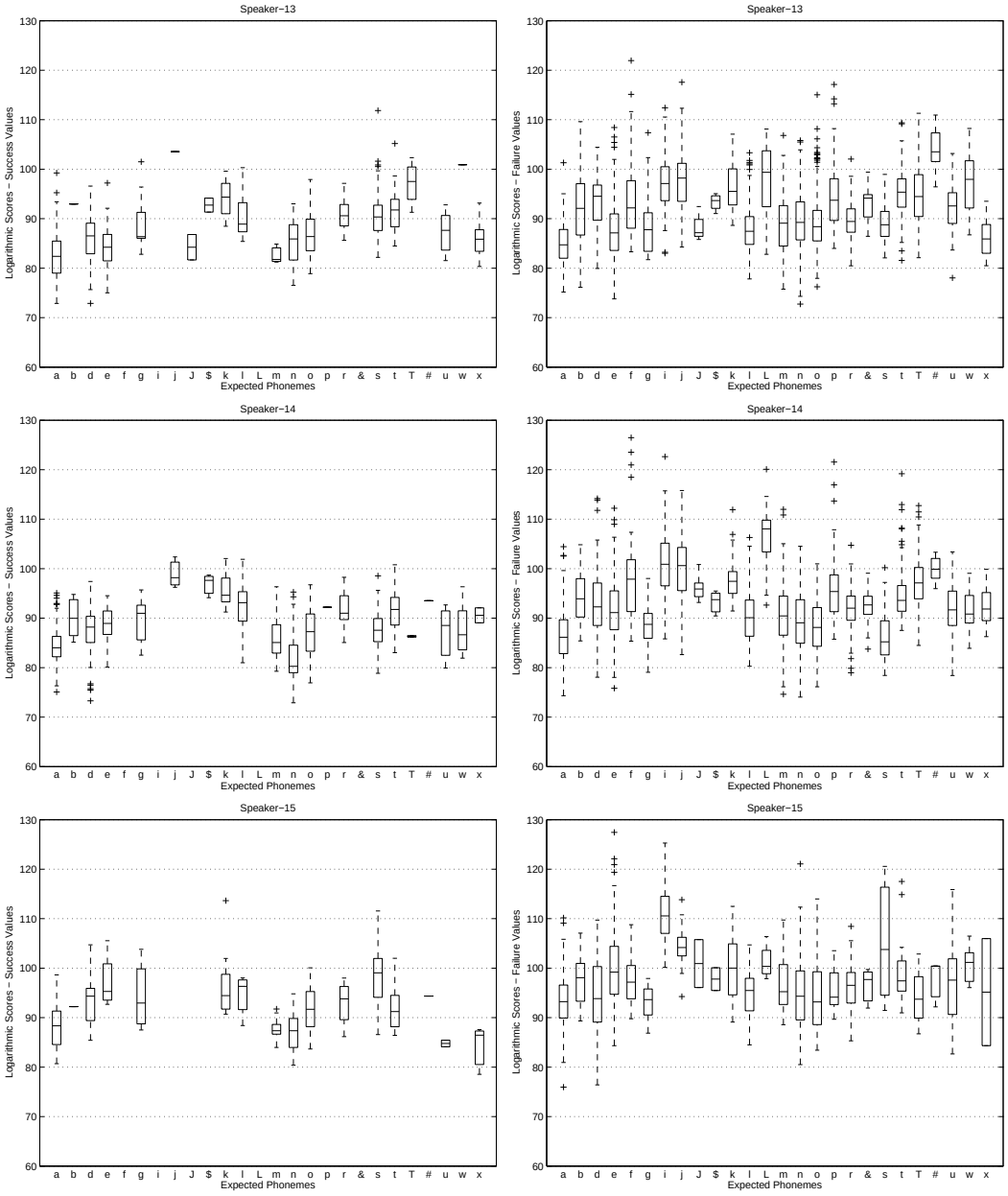


Figure 5.13: **Success and Failure logarithmic scores** for every phoneme. Boxplots comparison. Speakers f13, f14, L01.

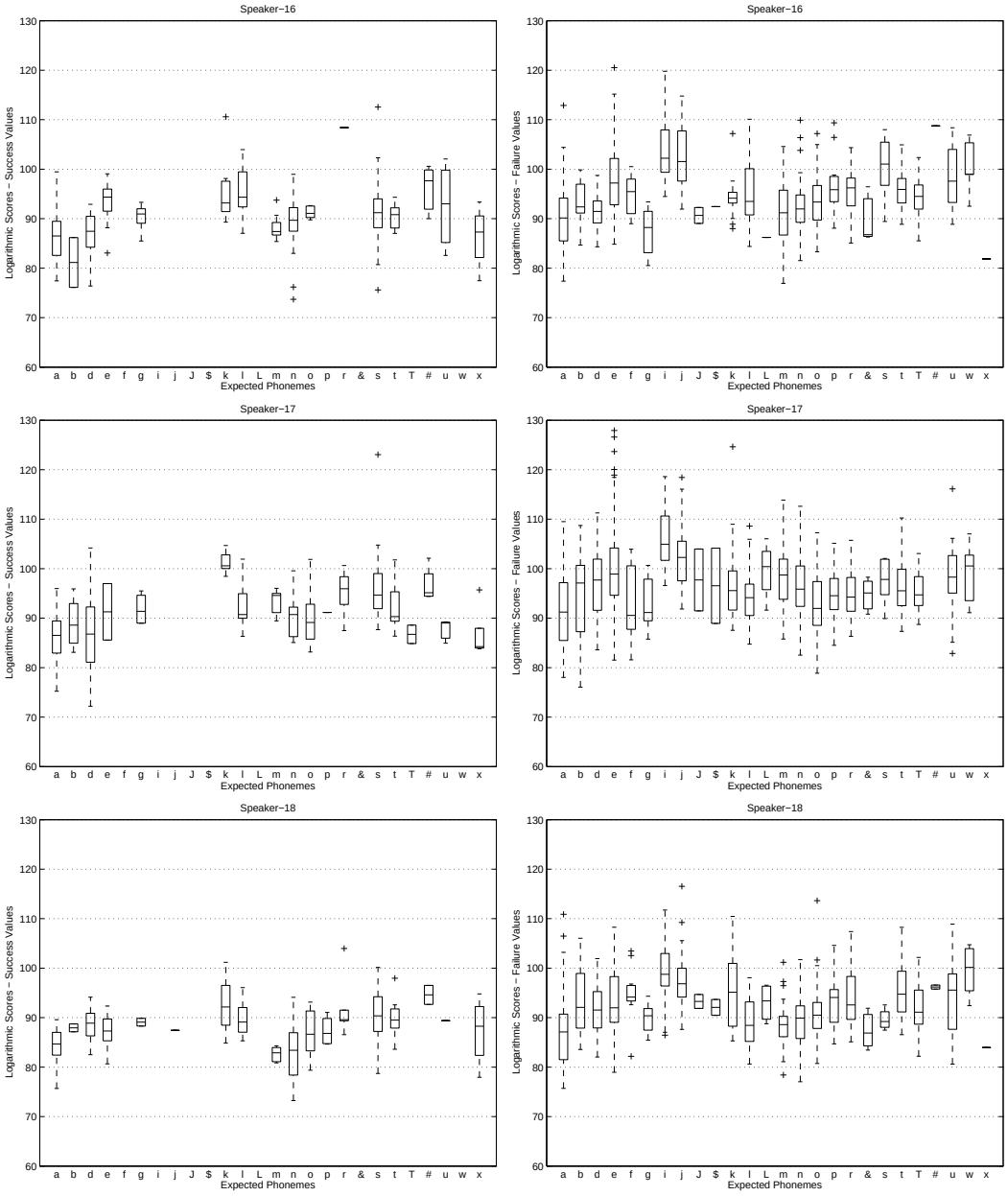


Figure 5.14: **Success and Failure logarithmic scores** for every phoneme. Boxplots comparison. Speakers L02, L03, L04.

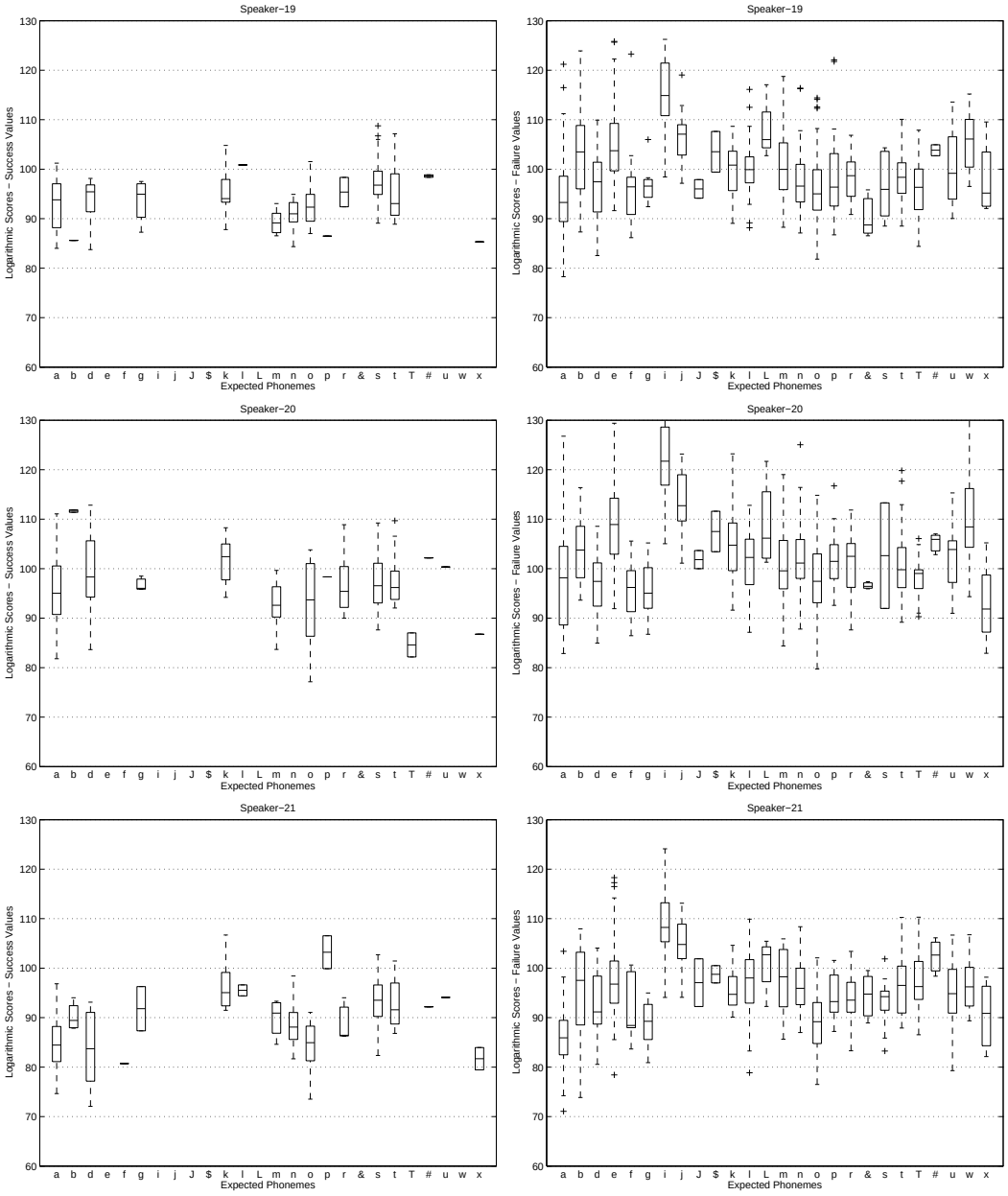


Figure 5.15: **Success and Failure logarithmic scores** for every phoneme. Boxplots comparison. Speakers L05, L06, L07.

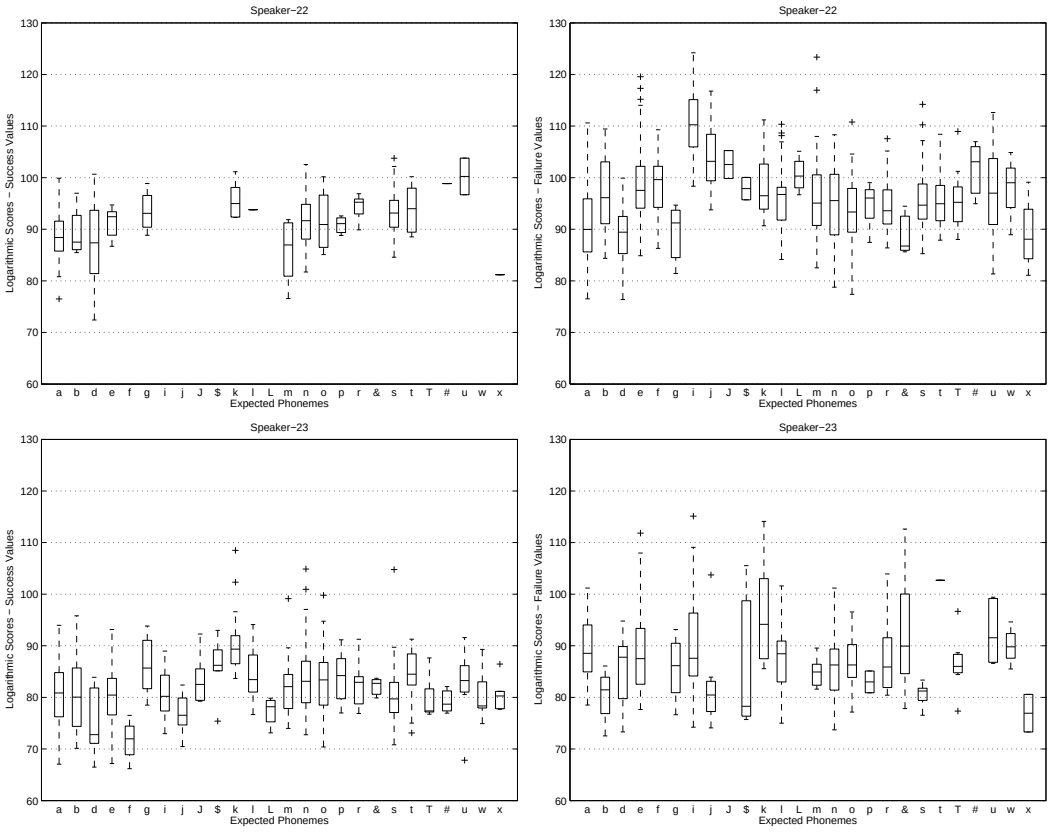


Figure 5.16: **Success and Failure logarithmic scores** for every phoneme. Boxplots comparison. Speakers L08 and Albayzin.

Figure 5.17: Occurrence of a detected phoneme for every expected phoneme. Non-Native Speakers (f01, f02, m03, f04, f05, f06).

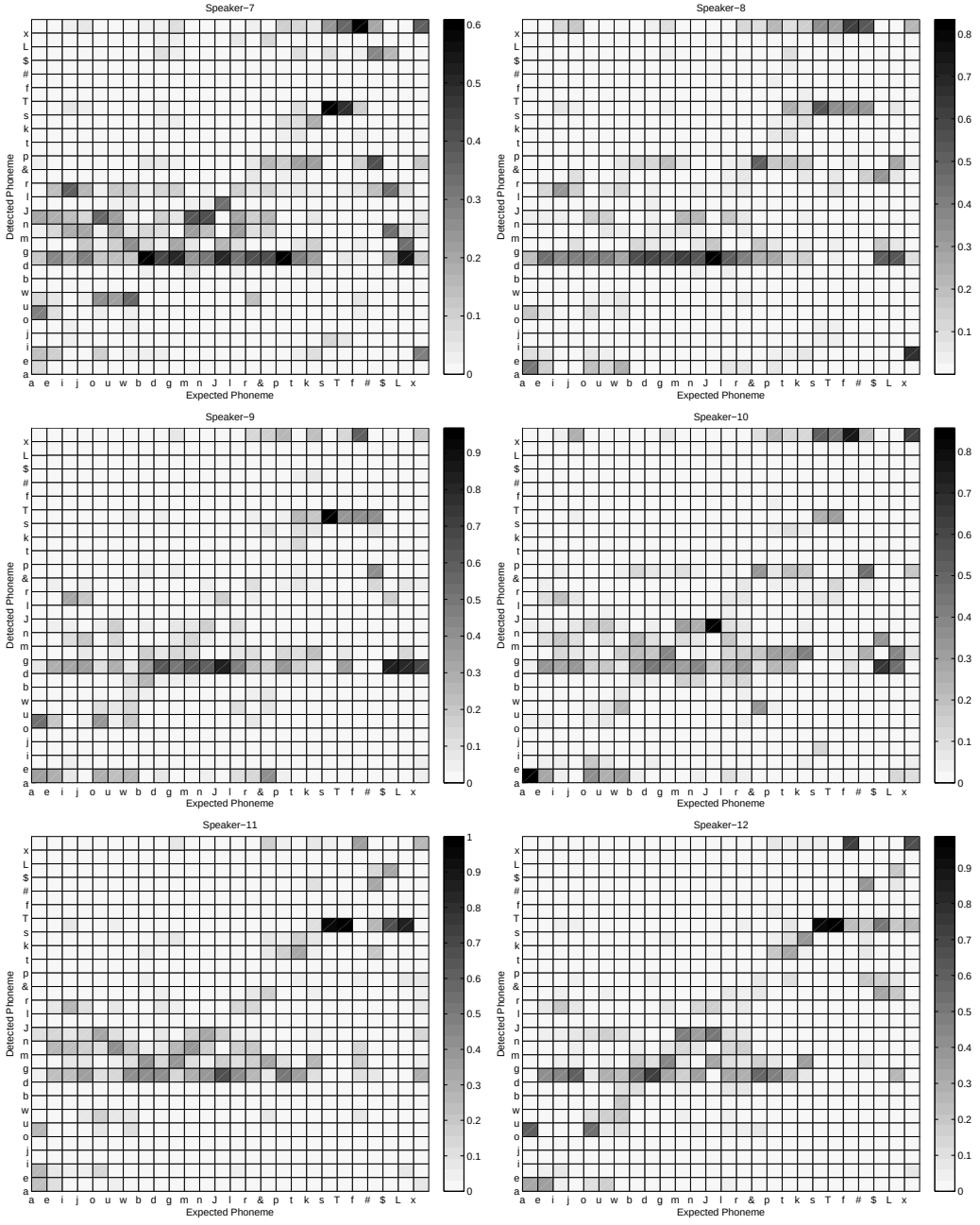


Figure 5.18: Occurrence of a detected phoneme for every expected phoneme. Non-Native Speakers (f07, m08, f09, f10, f11, f12)

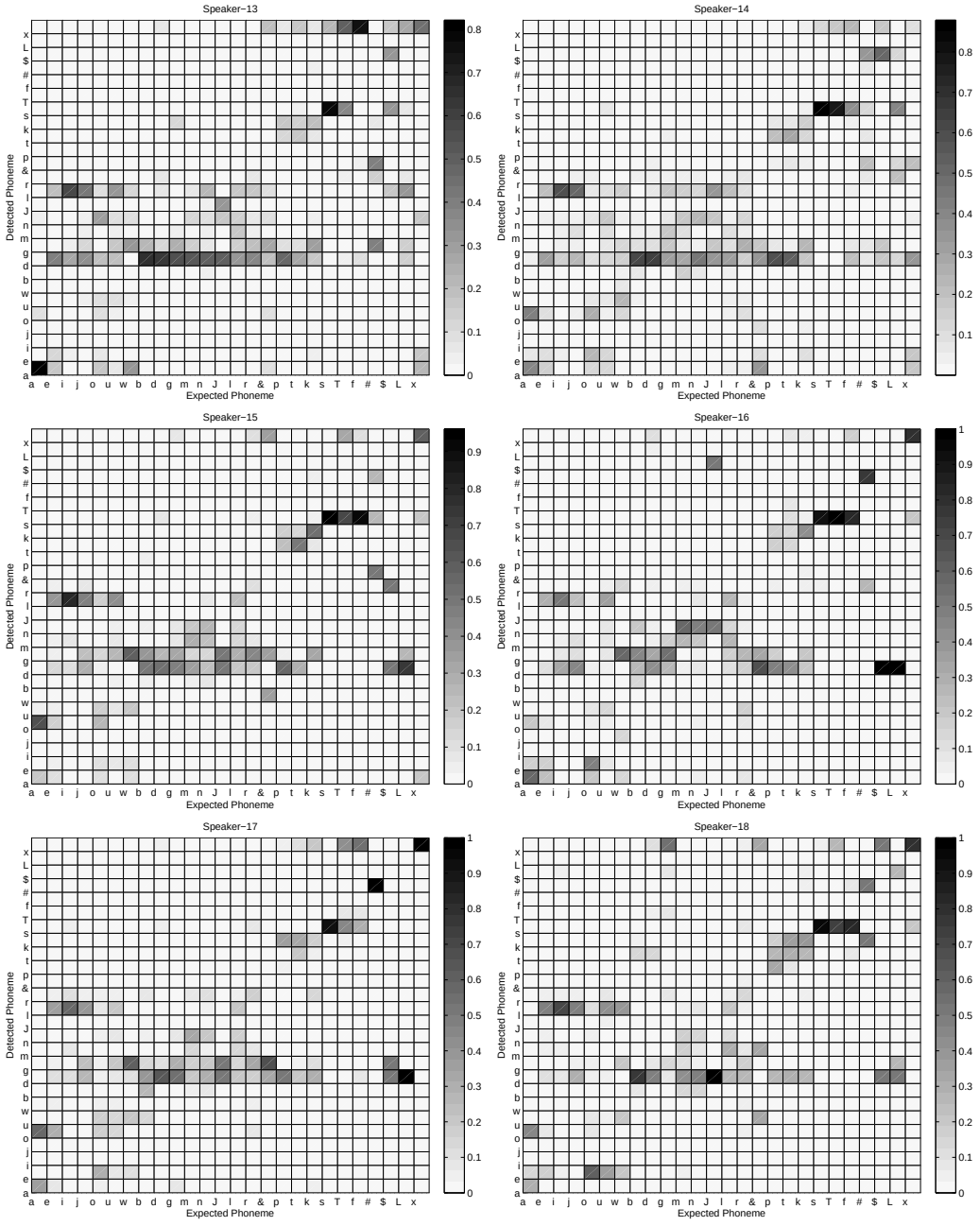


Figure 5.19: Occurrence of a detected phoneme for every expected phoneme. Non-Natives (f13, f14) and Natives (L01, L02, L03, L04).

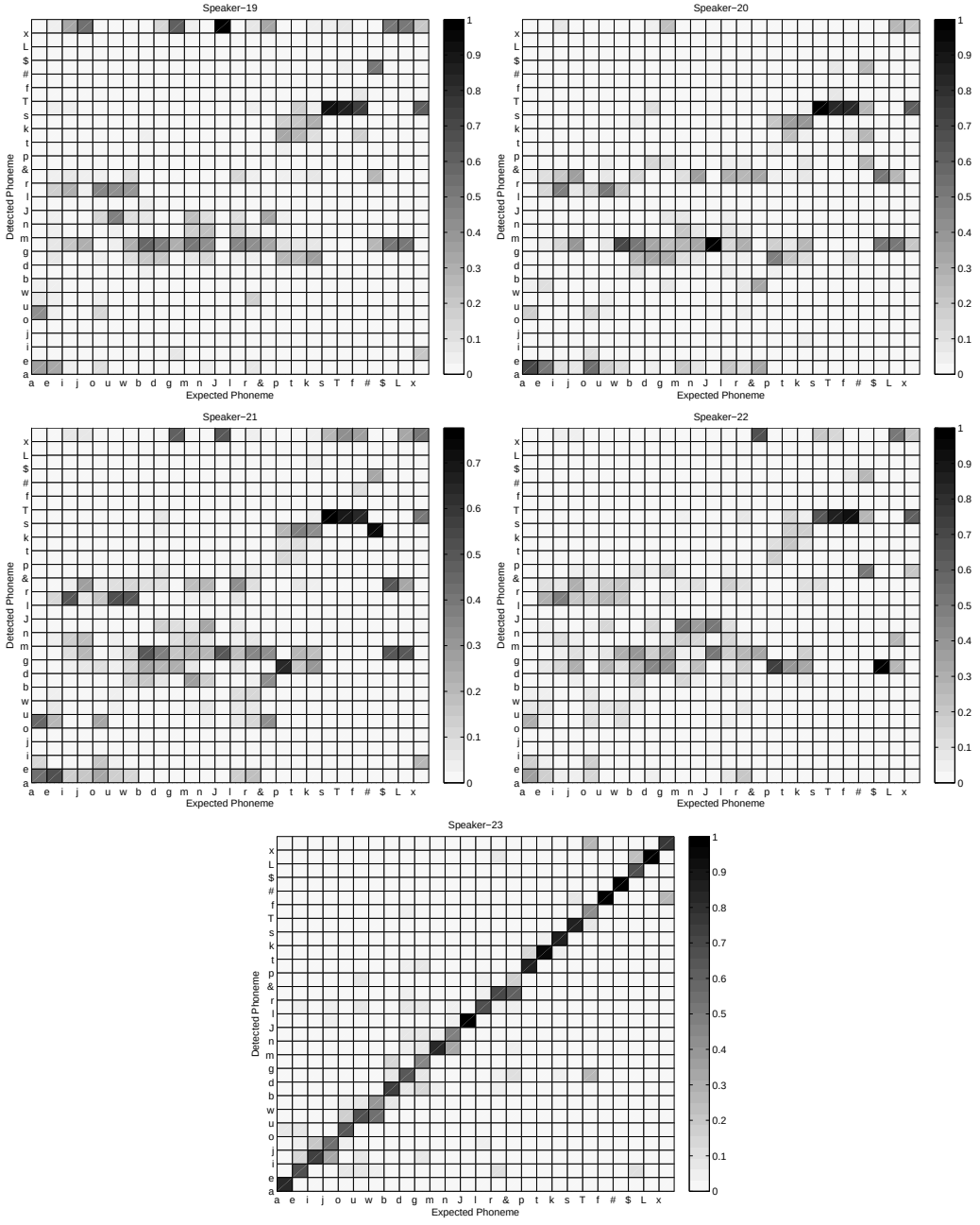


Figure 5.20: Occurrence of a detected phoneme for every expected phoneme. Natives (L05, L06, L07, L08) and Albayzin.