Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely accepted trials

Joaquín González-Rodríguez(1)*, Juana Gil*(2)*, Rubén Pérez*(3)*, Javier Franco-Pedroso*(1)

(1) ATVS-Biometric Recognition Group
Universidad Autónoma de Madrid, Spain
(2) Laboratorio de Fonética
Consejo Superior de Investigaciones Científicas, Spain
(3) Universidad del País Vasco, Spain
joaquin.gonzalez@uam.es

Abstract

Speaker comparison, as stressed by the current NIST i-vector Machine Learning Challenge where the speech signals are not available, can be effectively performed through pattern recognition algorithms comparing compact representations of the speaker identity information in a given utterance. However, this i-vector representation ignores relevant segmental (non-cepstral) and supra-segmental speaker information present in the original speech signal that could significantly improve the decision making process. In order to confirm this hypothesis in the context of NIST SRE trials, two experienced phoneticians have performed a detailed perceptual and instrumental analysis of 18 i-vector-based falsely accepted trials from NIST HASR 2010 and SRE 2010 trying to find noticeable differences between the two utterances in each given trial. Remarkable differences were obtained in all trials under detailed analysis, where combinations of observed differences vary for every trial as expected, showing specific significant differences in voice quality (creakiness, breathiness, etc.), rhythmic and tonal features, and pronunciation patterns, some of them compatible with possible variations across recording sessions and others highly incompatible with the same speaker hypothesis. The results of this analysis suggest the interest in developing banks of non-cepstral segmental and supra-segmental attribute detectors, imitating some of the trained abilities of a non-native phonetician. Those detectors can contribute in a bottom-up decision approach to speaker recognition and provide descriptive information of the different contributions to identity in a given speaker comparison.

1. Introduction and motivation

The speaker information present in the speech signal is spread across different information levels, ranging from low-level short-term acoustic and spectral features to high-level longer-term features [1]. There have been different successful approaches to integrate non-cepstral higher-level knowledge in speaker recognition systems in well-known NIST SRE tasks [1][2][3], but the complexity and limited increase of performance associated to using NIST-task-like higher-level systems results in cepstral-only systems being used in most (or all) practical scenarios.

However, non-cepstral information as voice quality features or particular pronunciation patterns could also be used in different ways. Current signal processing techniques and speech recognition (and understanding) systems provide an extraordinary ability to automatically detect acoustic and linguistic events in the speech signals. Once detected, those segmental and suprasegmental events can be further analyzed by specific detectors with ad-hoc feature extractors providing a rich picture of the non-cepstral characteristics observed in the given utterance. Those characteristics, as e.g. particular realizations of sounds, specific temporal patterns, degree of nasality, default phonation settings or particular voice quality in given linguistic segments, could act as soft biometrics for the speech signal [4]. Soft biometrics are characteristics or features of the individual that lack the distinctiveness to differentiate from any other person, but can contribute to filter or improve the decision process. For instance, in face biometrics, the eye colour or a given particular nose shape do not identify at all the person, but if two different persons have different eye colour or clearly different nose shapes it seems unreasonable to identify them as being the same person even if a state-of-the-art face matcher elicits a score higher than the acceptance threshold. A similar situation could be observed with i-vector systems in NIST SRE tasks, where very low false acceptance rates can be obtained, but with millions of trials we still face hundreds or thousands of falsely accepted trials where significant speaker information available in the speech signals under comparison is being ignored.

The information-extraction approach proposed by C.H. Lee et al. in [5] is of special relevance to this work, exemplifying a formal framework for a bottom-up attribute detection and knowledge integration in the complex task of robust speech recognition. Inspired by the processes involved in human speech recognition, that framework integrates non-synchronous evidences obtained at different segmental and suprasegmental levels, where the attributes found can serve as acoustic landmarks (also called “islands of reliability”). This modular strategy with expert-independent detectors shows several advantages, as the quality of each detector can be independently improved, and the overall detector combination procedure does not depend on the inner components of every detector. Moreover, similarities found from specific attribute detectors also have the advantage of providing descriptive useful information to the final user, far different from a single global score summarizing all the speaker-related information in the two utterances. In fact, our understanding about when and why in a given speaker comparison the system performs properly or not is very limited. In unknown scenarios, where reference data for calibration is not available, the elicited log-likelihood-ratio global score can be misleading, and objective and understandable descriptors can be extremely useful.

The experimental part of this paper tries to answer the question about the usefulness of those segmental and suprasegmental detectors in the audio conditions of a well-balanced and data-rich tasks as NIST SRE, where i-vector-based systems are performing extremely well. If this perceptual analysis shows to be useful in this task, it would provide a solid motivation to develop and integrate those types of non-cepstral detectors in our systems and decision-making processes.

The paper is organized as follows. After this introduction and motivation, in section 2 the contributions for speaker identity apart from cepstral information are reviewed. Section 3 describes how the experiments were performed, and Section 4 describes the results of the analysis. Section 5 summarizes the contributions from a phonetic point of view, and we express final conclusions in section 6.

2. Non-cepstral speaker information in speech

2.1. The multi-layered nature of speech

Every utterance made up of sounds is also emitted with a series of properties that are easy to be detected even by the untrained listener. This is so because in every speech act two well defined layers are superimposed: the verbal one (consisting of sequences of vowels and consonants) and the nonverbal one (which is conformed by a series of very different features –even nonlinguistic vocalizations, like coughs, snorts or throat clearing- that qualify and modulate the message). The ability to recognize familiar and unfamiliar voices that all listeners possess is based on the combination of all of these features. The most salient properties in the nonverbal layer (the so-called “high-level features” [1]), are suprasegmental, and they, in turn, could be sorted out in two major groups: those deriving from the different voice quality that speech can be produced with, and those referring to the prosodic level.

2.2. Suprasegmental features: Voice quality and prosodic features

All speakers have an intuitive knowledge of what, in phonetics, is called *voice quality* [6], and in non-specialist terms could be paraphrased as “the way in which people sound", the perceptual effect of their voices. The traditional definition of voice quality refers to "those characteristics which are present more or less *all the time* while a person is talking: it is a quasi-permanent quality running through all the sounds that issue from his mouth" [7]. This term is ambiguous, since, depending on the authors, it can refer to the permanent features of the person’s voice derived from both the phonation as well as the articulation of sounds or confined to the phonation type (modal, creak, falsetto, etc.) in which a specific statement is pronounced. In this paper, the term is always used in its first and broader sense, and as such, is understood to be the result of a series of articulatory settings and long-term modes of phonation conditioned by three types of factors:

a. Biological factors beyond the control of the speaker and of an extra-linguistic nature ("intrinsic", in Laver’s terminology [8]): the gender of the speaker, their age, physical attributes, and their state of health and medical history which largely determines their vocal attributes, such that some aspects of the sound of their voice are conditioned by the development and properties of their anatomy and the physiology of their vocal tract, whether they be permanent (i.e., cleft palate), semi-permanent (i.e., a particular hormonal state) or transient conditions (i.e., a cold). Two additional examples of the physiology of speaker-dependent traits are the upper and lower limits of their fundamental pitch range.

b. Social factors: those included in the language, dialect, or the sociolect to which the speaker belongs. Upon acquiring a language or a dialect variant, either by identifying them with a particular social group or family members, the speaker learns, by internalizing imitations, a series of articulatory/phonatory settings characteristics of that social collective which affect their quality of voice and mark their membership to that group [9].

c. Psychological and paralinguistic factors: personality traits, mood states or permanent or transitory emotions involving changes in voice [10][11]. Experiments performed in this respect confirm what the vocal correlates of different emotions are, and have provided fairly consistent results [12].

As regards prosody, it covers all those properties related to variations in pitch, intensity, and duration used to pronounce an utterance; that is, intonation (or tonal levels and contours), stress, accent, rhythm, and the speech rate (*speaking tempo*).

2.3. The multi-functional role of suprasegmental features

The aforementioned suprasegmental properties perform varied tasks in speech: they can denote grammatical meanings by differentiating the meaningful content of messages; they can provide a certain emotional or affective qualities to utterances; and when they are determined by the speech habits of the speaker or his anatomical features, they can also provide extra-linguistic information about their identity, as long as they are not sporadic or momentary and characterize complete fragments of speech [13].

Therefore, before considering the differences observed in the comparison of voices with respect to these suprasegmental variables to be idiosyncratic characteristics of the speaker, it is necessary to discern, firstly, if such differences are contingent on and motivated by grammatical reasons or by emotional and situational factors, all of which produce a great intraspeaker variability. For example, creaky voice is quite common in any language. The *creak* type of phonation can be easily used in many languages for linguistic or phonological purposes [14]: for instance, in some African languages there is a contrast between consonants pronounced with creaky voice and plain consonants which implies changes in word meaning; creaky voice can also have discursive and/or semiotic value [15][16][17][18]; for example, in Finnish and other languages, using creaky voice at the end of an utterance may indicate that the speaking turn is transferred to the partner. The creaky voice can also impart paralinguistic or emotional functions [19] and indexical or social functions: for example, in the English from Edinburgh, creaky voice is typical for speakers of high social position [20]. Creaky voice may be also a constant and purely individual feature, as an independent idiosyncratic habit. Many voices that are informally called *sepulchral voices* or *gruff voices* are in reality a creaky phonation type [10].

The same thing may be said about other voice qualities. For instance, *falsetto*, *breathy voice*, and *vocal tremor* are also relatively common voice qualities. *Falsetto* is not usually present as a stable characteristic except when used in a conscious way to create, for example, a certain sexual identity [21]. However, sporadic *falsetto* does appear in the speech of many people and is usually used to achieve greater expressiveness in conversational situations. The *breathy* quality, on the other hand, is extremely common in female speech. In [22] it has already noted that some women tend to phonate systematically with incomplete vocal cord closure in each phonatory cycle, which produces a greater release of air and an increase of concomitant noise, and this gives rise to a *breathier* quality. Finally, *vocal tremor* is a feature that, within normal limits, is manifested as a sporadic or momentary alteration in the control of the phonation mechanism (vibrato-like modulations), indicating a general deterioration as part of the natural aging process, or the appearance of certain diseases [23][24].

Voice qualities motivated more by articulatory settings than by peculiar phonation modes, such as pharyngalized voice or voice with lip protrusion, for example, can also perform an affective function (last one is frequent in mother-to-baby speech), or can be the result of a particular habit.

2.4. The discriminant performance of suprasegmental and segmental features

The individualizing capacity of voice quality and prosody is determined by its degree of permanence in the speech of the speaker and the degree in which he can exercise control over them. In this sense, there could be considered to be degrees of discriminatory capacity. For example, the use of a particular articulation rate is a feature that the speaker can control and alter depending on the context. Therefore, determining whether or not it is a unique permanent feature of a speaker from a single sample of speech is impossible, since a large number of samples in different contexts would be required.

A dysphonic voice, on one hand, is certainly beyond the control of the speaker, but as in the case of articulation rate, may not be the result of a permanent abnormal phonation of the speaker, but only a temporary consequence of a cold or temporary laryngitis. On the other hand, the *continued* presence of creakiness in a voice sample is not caused by grammatical, situational or emotional factors, nor is it easily controllable or alterable by the speaker, because it responds to a learned phonatory habit (whose roots could be sociodialectal or idiolectal) and therefore is valuable for distinguishing between individuals. In the analyzed pairs of voices in this work, this level of discriminatory potential makes the presence of some discriminant features incompatible with the hypothesis that a single speaker is behind the two compared samples, while the presence of other features only raises doubts about whether there is one person or two.

The same could be said of the next group of properties that have been taken into account in this study, that is, the (non-cepstral) segmental features that differentiate each pair of voices, and which, due to their different status, can have varying degrees of differentiating power. More or less reduced vowel pronunciation, for example, is concomitant with the articulation rate in a way that, if the rate is altered, the reduction is likely to be attenuated or increased proportionally. On the other hand, a unique and essentially unchanging articulation, in any syllabic context, of the fricative /s/, for example, is easy-to-perceive by any listener and is an extreme differentiator. The performance of various properties, therefore, is also different at this verbal level.

3. Description of the experiment

The co-authors team of experienced phoneticians accepted to perform, being a heavy time consuming task, an in-depth analysis of some i-vector-based falsely accepted trials from NIST 2010 HASR (Human Assisted Speaker Recognition) and SRE 2010 telephone-only tasks. Those phoneticians knew in advance that in all cases the speakers in every trial were actually different. However, they are not asked to perform speaker identification but just to detect differences between the two recordings in every trial. If those differences are compatible or not with the same speaker hypothesis is out of the scope of this paper, our interest being focused in knowing if a well-trained phonetician can detect differences beyond the extremely similar spectral characteristics detected by the i-vector systems which erroneously favored a same speaker hypothesis based only in the cumulative short-term spectral information.

3.1. Selection of trials

The submitted scores from ATVS-UAM i-vector system at SRE 2010 [25] were used for the pre-selection of trials. First of all, 16 impostor trials with submitted calibrated likelihood ratios favoring the same speaker hypothesis (logLR>0) were selected from the HASR trial list. Additionally, as HASR trials were manually selected by the organizers to be perceptually very similar, an extra set of 50 regular trials was extracted from the tel-tel SRE10 task. In this case, impostor trials were selected among those with larger errors (strong support to the wrong hypothesis) by our i-vector system, resulting in those in the 3<logLR<5 interval.

Table 1: List of NIST SRE10/HASR trials under analysis

Target ID	Test utt.	Gender	Task
30853	txhnh_b	f	HASR
30973	txbtg_b	m	HASR
31204	tbrpz_b	f	HASR
31925	tdmsx_b	f	HASR
32891	tufqo_a	m	SRE10
32906	tfafy_b	m	SRE10
32971	tgkig_b	m	SRE10
32986	tyopp_a	m	SRE10
33508	tphpl_a	f	SRE10
33785	tjmiz_b	m	SRE10
34078	tqkbc_b	f	SRE10
34245	tlbuw_a	f	SRE10
34394	tphjx_b	f	SRE10
34412	tgwsc_a	m	SRE10
34819	tejub_a	m	SRE10
34965	tsaoq_a	f	SRE10
35172	tjqmc_b	m	SRE10
35257	tqrxi_a	f	SRE10

The voices that made up those 66 pairs were extremely similar, and were obtained through semi-directed talks by the corresponding speakers with their conversation partners whose voices can be heard in some cases, but not always. The recordings were made in very different conditions: by phone, direct recordings with noise background, etc. The speakers are men or women of apparently different ages, and all English speakers. All examined conversations are relaxed and revolved around general topics. Although all sound samples last about five minutes, the amount of speech from the informants differs due to the nature of the talks and/or the communicative interest of the interviewer or interlocutor.

The task of the phoneticians involved in the study - who knew in advance that the voices did not belong to the same speaker despite their similarities – consisted, during the initial phase, of specifying whether they could find perceptual differences between the two voices in each pair, then check if the observed differences were corroborated by the acoustic analyses of the samples. After a quick first listening (approx. 10 min.) of each of the 66 pairs, 18 of them were selected, 9 male voices and 9 female, 4 belonging to NIST HASR 2010 and 14 to SRE 2010 (about 25% of the selected trials in both cases – 4/16 in HASR and 14/50 in SRE10), because they presented immediately noticeable differences in the opinion of both phoneticians, and therefore were more susceptible to analysis. The selection criterion was, consequently, the appreciation of differences between the two recordings from the first moment.

3.2. Perceptual analysis

In the second phase, the two phoneticians conducted a more detailed perceptual analysis, for which they prepared a list of potentially defining characteristics of the voices (see left column in Table 2) from the suggestions listed in the bibliography (cf. [26], p. 179). None of the already existing protocols were used for the perceptual analysis because almost all of them are designed for specific purposes –usually medical- and pay more attention to the variables related to modes of pathological phonation, for example, than the variables that are non-pathological and more properly linguistic, such as those referring to syllabic structure, or prosody. Both phoneticians heard, together and on several other occasions, each of the pairs of analyzed voices, and noted on the features list which, in their judgment, distinguished each voice from its counterpart.

Subsequently, each of the two analysts independently listened again to the constituent voices of each pair in order to check if the features observed in the first listening continued to be perceived, and to discard any hint of mutual influence in their evaluations. Finally, the recordings were acoustically analyzed (intensity, first and upper harmonics amplitude, mean value of the f_0, jitter, shimmer, HNR –harmonic to noise ratio-, etc.) using the PRAAT program [27], to confirm if the phoneticians' auditory impressions could be corroborated by objective quantifiable data.

It should be noted, however, that the acoustic analysis is not always sufficient to characterize a certain voice quality. Some phonatory registers present a considerable overlap in various acoustic dimensions, as explained in [28]. It is the characteristic physiological aspect of each type of phonation or of each articulatory setting that ultimately distinguishes it from others, even though the correspondence with the combination of specific acoustic features may vary. Hence, in a context beyond the scope of this work, an acoustic-perceptive study of any voice quality type would have to be completed at a later stage with one that includes physiological, articulatory and aerodynamic analysis.

The detailed list of the trials under analysis is shown in table 1. During our following discussion of results, trials are identified by the first number in the trial description (first column in table 1). Perceptual and acoustic analysis of each pair, jointly and individually, took about one hour altogether. Detailed reports documenting every analyzed trial, including precise descriptions of the instants of time where features were found with the results of the corresponding instrumental analysis for each feature are available, but for obvious reasons of space are not included in this paper.

4. Results and Discussion

4.1. A difficult but achievable task

The task of comparing the selected voice pairs doubtlessly presented much difficulty because they were very similar, trials being selected from falsely accepted trials by the i-vector system. From the perceptual (and later acoustic) analysis performed, it is observed in the first place that, although these similarities are unquestionable, it is also possible to perceive marked differences among the voices - very noticeable in some cases. Those divergences are related to one or several unshared ('high-level') features, basically prosodic and voice quality features, and, to a much lesser extent, segmental features, and are summarized in table 2. In the authors' experience, we expect that a more comprehensive and extended study of the pairs of voices that were initially disregarded for the detailed analysis in this work would also provide results in the same direction. It is true that the analysis was performed over the 25% of perceptually "easiest" trials from the original selection of 66 i-vector-based falsely accepted trials. However, the authors hypothesize, from their first short informal hearing, that they could have equally found differences in most (or all?) of the cases, probably devoting more (or much more) than one hour of analysis per trial as they needed for the "easy" 18 trial set.

4.2. Voice quality as a prevailing individualizing feature

4.2.1 Phonation types

Table 2 clearly shows that the features most often perceived as aurally different relate to the quality of voice, in particular, to the phonatory settings that generated them (in 14 out of the 18 pairs we found clear differences in the phonation type).

This is initially coincident with what -according to the experience of the signing phoneticians of this study- often happens in abundance for forensic purposes: when there is a type of phonation different from the modal in one of the compared samples, the potential for discrimination between two voices increases significantly. Often it is harder to perceive the effects of a voice quality motivated only by supralaryngeal settings (as it was observed in trial 32891, wherein one of the voices appears to respond to a certain degree of lip rounding, non-existent in the other). This is plausible, because the latitudinal or longitudinal adjustments of the vocal tract that can alter the voice quality usually do not have the same degree of perceptual effect as changes in phonation type; very often they nuance it, but don't change it drastically.

At trial 30853, for example, one of the voices is dysphonic (whether permanently or temporarily is not known), and, at the same time, shows some anomalous features that suggest a peculiar overall articulation. In fact, dysphonia was immediately identified; on the other hand, the special voice quality that accompanied it was very difficult to determine: it could be due to excessive salivation, to a special arrangement of teeth or a special configuration of the palate, or several causes at the same time. This voice, also, seems to have an excessive damp secretion in the respiratory tract and it is the only one among those analyzed in which hyponasality was observed, perhaps concomitant with the dysphonia (may be due to a cold).

4.2.2. Creaky voice

If a subsequent comparison is made of the various phonation types in the given set of trials (Figure 1), the *creaky voice* quality of one of the two collated voices is the most frequently documented mode of phonation, followed by dysphonic emissions and other distinctive types of phonation (vocal tremor / vibrato, falsetto, breathy).

Episodes of creaky voice are quite common in English (mostly in the North American variety), as has been pointed out before by many authors ([22][29], among others) and all conversations analyzed for this study are in English. It was expected, then, to find creaky voice. What is not so common is to find that all utterances of a speaker are issued with creaky voice, as it happens in trial 30973 or in trial 32096, in which virtually all of the emissions produced by the voice 2 of each one of them are made with creak, unlike what occurs in voice 1, which appears modal. In the voice 1 of 30973, the speaker lowers the pitch and, possibly, the larynx, especially at the end of his statements, but is almost never creaky. Such a phenomenon does not respond to specific constraints of any kind (grammar, emotional or otherwise), but it is an acquired and distinctive habit of the speaker.

4.2.3. Other phonation types

The other phonation types appear much less often. Trial 34394 is particularly interesting as the second female voice, very similar to the first one, presents, however, air leaks during all her emissions, pointing to a breathy phonation. It is difficult to find this kind of setting, with a lot of aspiration noise, on male voices, and it is easy to recognize on female ones in all languages. This is a mode of phonation that transmits intimacy or -in cases of very intense aspiration- sensuality, but it is always lack of tension. That explains why the intensity range of the second voice of the 34394 trial is much lower than that of the first voice, and maintained at a low level, which was also very noticeable. Also, episodes of falsetto voice and tremor or vibrato are present respectively in trials 31925 and 34245, but remain only episodic occurrences. In the latter case, the voice was particularly strident, and with a very slow tempo, so the speaker could be an old person, which would explain the sporadic appearance of tremor.

4.3. Time domain features

In addition to the quality of voice, the speech rate (*tempo*) as well as inclusion of pauses and other timing factors have clearly proven to be distinguishing features as previously mentioned in the literature [30][31][32][33]. The speech rate in general (with pauses included) and articulation rate (without pauses) can be important indications of many characteristics of the speaker, as are age, their language, and tasks performed, i.e., reading or conversation.

Table 2: Summary of differences found across the speakers in every trial

Trial	30853	30973	31204	31925	32891	32906	32971	32986	33508	33785	34078	34245	34394	34412	34819	34965	35172	35257	#
Pitch																			
F0 mean											X				X				2
Pitch range																	X		1
Loudness																			
Loudness variability	X	X		X	X						X	X	X					X	8
Phonation type																			
Dysphonia	X													X				X	3
Creaky		X	X		X	X		X	X			X					X		8
Breathy													X						1
Falsetto				X															1
Tremor / Vibrato												X				X			2
Temporal characteristics																			
Speech rate	X			X	X					X		X						X	6
Pauses					X														1
Length of breath groups					X														1
Vowels and consonants duration					X														1
Articulatory settings																			
Lip protrusion					X														1
Articulatory characteristics																			
Hyperarticulation	X			X	X							X							4
Coarticulation		X	X		X														3
Vowels and consonants realization	X		X		X														3
Degree of nasality																			
Hyponasality	X																		1
Prosodic line																			
Falling / Rising						X													2
Tonal contours				X			X			X				X					4
Hypermelodic/Monotonous							X				X	X	X					X	5
Syllabic structure																			
Full				X															1
Other features																			
Dialect / Idiolect / Accent				X						X				X					3
Specific voice impression (strident, muffled, wet, …)	X			X				X			X	X				X	X	X	8

Considering, for example, the trial 32891 (whose first voice was already mentioned when talking about the lip protrusion), between the first and the second voice, the huge difference in the temporal parameters draws attention: once the conversational contexts are equivalent, the second speaker articulates very slowly, sometimes even with difficulty, as if he couldn't find the words he wants to use (in certain very specific fragments his speech reminds of that produced by intoxication). He prolongs certain sounds, and his overall production is faltering. The temporal features are, in fact, very important for the discrimination of voices. It may be noted in this regard that when a speaker tries to disguise his voice -changing, for example, his phonation mode-, a feature that has proven key to being recognized by listeners is the speech rate [34][35]. Usually, the rate at which a person speaks corresponds to the degree of hyper- or hypo-articulation and with the degree of co-articulation among sounds, so that the second voice sounds in trial 32891 are not reduced or easily elided, that is, the intelligibility is greater.

4.4. Intonation

The mean value of the fundamental frequency, the range in which it moves, and the stress and tonal patterns that constitute intonation, as well as the general melodic line -falling or rising- with which the utterances are modulated were elements that helped differentiate several trials. The f_0 is always an important cue in almost every study of voice, because it is also very robust perceptually against environmental noise. It is one of the first features that phoneticians compare in abundance. In this case it served, for example, to clearly distinguish between the voices of trials 34078 (with average frequencies of 184 Hz vs. 131 Hz in female voices) and 34819 (male voices with 140 Hz vs. 65 Hz). As expected, given the presence of creak, mean frequency is very low. For example, creaky voice 2 in trial 30973 has a mean frequency of 86.26 Hz (values of a creaky voice tend to range between 70 Hz and 80 Hz), its standard deviation is 11.70 Hz, the higher f_0 is 124.94 Hz and the minimum is 68.25 Hz. If these values are compared to those characterizing the voice 1 in a correspondent fragment of the recording, also declarative and therefore comparable, the obtained values are different, corroborating previous perceptual impressions: a mean frequency of 121.37 Hz; a standard deviation of 31.08 Hz (i.e., the variation around the mean is much higher, which is what modal voice permits); a highest pitch of 278.32 Hz and a minimum of 55.37 Hz, or, in other words, a much wider pitch range, which is also expected.

Interesting speaker idiosyncrasies, capable of distinguishing some of the speaker pairs, were also found on the level of accentual and intonational patterns (for instance, in trials 31295 and 32971), which requires some in-depth study beyond the scope of this paper. In any case, in [36] the relevance of these factors for speaker recognition was proved by means of identification and discrimination tests with eight listeners, confirming what was noted previously in [37]: f_0 mean and pitch patterns provide extremely valuable information about the speaker.

4.5. Segmental features

Among the analyzed pairs some were found which differed because one of the voices presented particular realizations of certain sounds, namely the fricative /s/. This happened in the trial 31204. Regarding the first voice of this pair, the sound of this fricative does not seem to correspond, in most of its occurrences, to an apical-alveolar strident articulation, common in English, but rather to a mellow realization and laminal-dentoalveolar fricative type. It was found in the acoustic analysis that its energy seems reinforced around 2500 Hz, and that there is a great general inharmonicity over all frequencies. In the second voice, instead, the start may be located around 3000 Hz, with most intense peaks and a lower dispersion along frequencies. This means that the /s/ voice of the first voice is less strident and his acoustic characterization approaches that of /θ/, as shown perceptually.

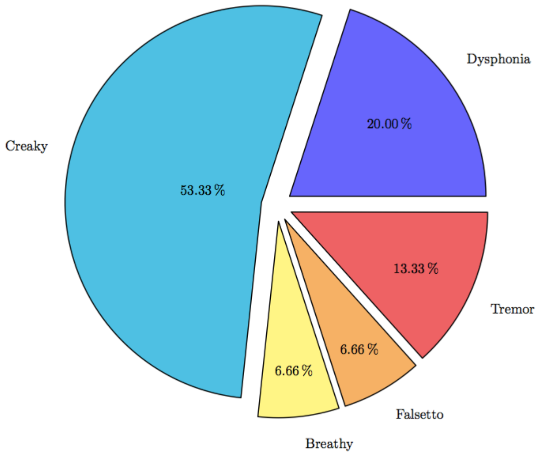

Figure 1: Percentage of total number of differences in phonation types found

It is possible that the anatomy of the oral cavity of this speaker shows some peculiar trait, or perhaps her particular pronunciation of /s/ is simply due to an unusual learning pattern. In any case, the perceptual impression produced by her articulation of another fricative, the /f/, is also characteristic, which supports the hypothesis that there may be underlying anatomic causes in her idiosyncrasies. Interestingly, this difference between the two realizations of /s/, with their corresponding distinct perception, also occurred in trial 30853, which was mentioned above when discussing the emergence of dysphonia in one of the voices.

In this very trial 30853 further details could be seen relating to the realization of segmental elements. The pronunciation of some words that appear in the recordings of the two voices is different. For example, the word '*here*' appears twice in each of them, and does not sound alike. Once the acoustic analysis was carried out, it was found that the formants of the diphthongized vowel [ɪə], which constitutes its nucleus, differ significantly in their values. In the first voice, values are prominently centralized, with a slightly higher F1 (more open vowel) and a significantly lower F2 (more central vowel), which begins in 2200 Hz and ends in 1600 Hz. There are no abrupt changes in the corresponding path from one vowel to another, as there is a high degree of co-articulation in this voice. In the second voice, on the other hand, more extreme values can be appreciated, with an initial more clearly closed vowel, whose F1 also starts around 300 Hz and whose F2 is considerably higher. This vowel is, therefore, a much peripheral vowel which requires more articulatory effort, (greater lingual displacement), and is maintained at least about 100 Hz higher than in the first voice throughout the completion of the diphthong. All this means that this second voice is more hyper-articulated, so it emphasizes the differences between the vocalic qualities.

4.6. Idiosyncratic features

Finally, there may be other idiosyncratic characteristics of the speakers not related either to types of phonation nor articulatory settings nor prosodic factors. In the case of trial 32986, for example, the two voices differed because the second speaker appears to have an altered respiratory function (noise breathing) throughout the entire recording, with occasional episodes of false notes (squawks). These features, which are rare, as any pathological or semi-pathological feature, are therefore very discriminatory.

5. Summary from a phonetic perspective

Several concluding remarks are inferred from the performed perceptual analysis:

a) Human perceptual capacity is able to detect features with high discriminatory power among compared voices, even though the degree of similarity among them is very strong, as long as careful and repeated listening is performed.

b) The properties that determine how a voice sounds are highly interrelated, both according to the systems of muscle activity underlying laryngeal and supralaryngeal systems that generate them, and according to the very process of perception: the perceptual effects of different features can be mutually enhanced or masked. Therefore, in the perceptual process, the detection of the presence of one of these properties can help to highlight other secondary cues associated with it that contribute to the final perception. For example, a lower speaking tempo very frequently implies a lesser degree of overlapping between adjacent sounds and a general impression of hyper-articulation.

c) The difficulty of determining the specific contribution of each of these elements to the overall result is derived from the above considerations, because, firstly, the final perception is multidimensional and clearly holistic or, if preferred, gestalt, i.e. the all you hear is more than the sum of the parts; and secondly, the perceptual importance of different elements of voice quality can vary depending on the time, the situation or the listener condition. Certain phonatory or articulatory information may be decisive in the discrimination and recognition of a certain voice and not in others (see [26] for a more detailed discussion).

d) All of the above being true, it seems obvious, because of the results on this study and the experience that the signing phoneticians have in their daily work, that the presence of types of phonation others than modal is a trait that, alone or associated with others, has a high discriminatory value. More generally, the features associated with the voice quality are found of fundamental importance for the discrimination of the speaker.

e) The prosodic features associated with the modulation of the fundamental frequency, either stress or tonal features, are also a very important element of discrimination. An idiosyncratic intonation pattern, a recurrent final tone, a syllabic or accentual rhythm can be key in discriminating one voice over others. They are learned features and therefore potentially controllable by the speaker, but they are usually so deeply learned habits that they are not easy to alter. Similarly, the tendency of speakers to move in a certain loudness and frequency range, their preference for a pitch (tessitura) and for a specified volume, influence the general perception of the voice like the other mentioned features.

f) The characteristics associated with timing (prolongation of sounds, extension of breath groups -that is, the portions of speech between pauses-, the frequency of interleaving silences, the average length of those silences, the particular features of sounds that sometimes serve to fill them, the general speech rate, etc.) are all elements that allow the extraction of information of the analyzed voice.

g) Finally, the particular realizations of certain sounds can be very valuable evidence to assess the compatibility of two voices with the hypothesis of a single speaker; the more valuable the more they deviate from the expected stereotype with which the listener unconsciously compares them.

6. Conclusions

Phoneticians have performed perceptual and instrumental analysis of voices in a wide variety of situations, but the potential of their approach in audio recordings from a well-formed and balanced task with low error rates as NIST SRE was largely unknown. In this work, our non-native team of phoneticians performed a detailed analysis over difficult speaker comparison trials, where significant non-cepstral segmental and supra-segmental differences among the speakers in each pair of utterances per trial were carefully documented.

The conclusions of this work are twofold: firstly, we have confirmed the usefulness of all the non-cepstral information that is currently ignored even in a task where an i-vector-system is performing extremely well. And secondly, the ability of non-native phoneticians to perform with limited effort such detections motivates research towards the development of banks of attribute detectors as voice quality features or particular pronunciation patterns, imitating some of the phoneticians' trained abilities, in order to provide objective and descriptive information that could contribute in a bottom-up information-extraction approach to speaker comparison.

7. References

[1] Shriberg, E., "Higher-level features in speaker recognition", In: C. Müller (Ed.) Speaker Classification, Vol I, pp. 241-259, Berlin: Springer Verlag, 2007.

[2] Shriberg, E., et al. "Modeling prosodic feature sequences for speaker recognition." *Speech Communication* 46.3 (2005): 455-472.

[3] Kockmann, M., Ferrer, L., Burget, L., Shriberg, E., & Cernocky, J. (2011, May), "Recent progress in prosodic speaker verification", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4556-4559, 2011.

[4] Park, U., & Jain, A. K., "Face matching and retrieval using soft biometrics", *IEEE Trans. on Information Forensics and Security*, 5(3), 406-415, 2010.

[5] Lee, C. H., & Siniscalchi, S. M., "An Information-Extraction Approach to Speech Processing: Analysis, Detection, Verification, and Recognition", *Proeedings of the IEEE*, 101(5), 1089-1115, May 2013.

[6] Laver, J., *The Phonetic Description of Voice Quality*, Cambridge: Cambridge University Press, 1980.

[7] Abercrombie, D., *Elements of General Phonetics*. Edimburgh: Edinburgh University Press, 1967.

[8] Laver, J. "The semiotic nature of phonetic data", *York Papers in Linguistics*, Vol. 6, pp. 55-62, 1976.

[9] Campbell, N. "Changes in voice quality due to social conditions", *Proc. of the XVI International Congress of Phonetic Sciences*, Saarbrücken, pp. 2093-2096, 2007.

[10] Laver, J., "Voice quality and indexical information" en *British Journal of Disorders of Communication*, Vol. 3, pp. 43-54, 1968.

[11] Laver, J. *Principles of Phonetics*, Cambridge: Cambridge University Press, 1994.

[12] Gobl, Ch. and Ní Chasaide, A. "The role of voice quality in communicating emotion, mood, and attitude", *Speech Communication*, Vol. 40, pp. 189-212, 2003.

[13] Nolan, F., French, P., McDougall, K., Stevens, L. and Hudson, T., "The role of voice quality 'settings' in perceived voice similarity, abstract presented at the *IAFPA Conference 2011*, Vienna, 2011.

[14] Gordon, M. and Ladefoged, P., "Phonation types: A cross-linguistic overview", *Journal of Phonetics*, Vol. 29, pp. 383-406, 2001.

[15] Ogden, R., "Non modal phonation and turn-taking in Finnish". In E. Couper-Kuhlen and C.E. Ford (Eds.) *Sound Patterns in Interaction: Cross-Linguistic Studies from Conversation*, Amsterdam: John Benjamins, pp. 29-62, 2004.

[16] Campbell, N. and Mokhtari, P., "Voice quality: The 4th prosodic dimension", In: *Proceedings of International Congress of Phonetic Sciences*, Barcelona, Spain, pp. 2417-2420, 2003.

[17] Bőhm, T. and Shattuck-Hufnagel, S. "Listeners recognize speakers' habitual utterance-final voice quality". In: *Proc. ParaLing2007*, pp. 29-34, Antwerp, 2007.

[18] Mendoza-Denton, N., "The semiotic hitchhiker's guide to creaky voice", *Journal of Linguistic Anthropology*, Vol. 21, n. 2, pp. 261-280, 2011.

[19] Johnstone, T. and Scherer, K. "The effect of emotions on voice quality". In: *Proc. of the XIVth International Congress of Phonetic Sciences*, San Francisco, USA, pp. 2029-2032, 1999.

[20] Esling, J., "The identification of features of voice quality in social groups*", Journal of the International Phonetic Association*, Vol. 8, n. 1-2, pp. 18-23, 1978.

[21] Podesva, R.J., "Phonation type as a stylistic variable: The use of falsetto in constructing a persona", *Journal of Sociolinguistics*, Vol. 11, n. 4, pp. 478-504, 2007.

[22] Henton, C. and Bladon, A. "Creak as a sociophonetic marker". In: V. Fromkin, L. M. Hyman, and C. N. Li (Eds.) *Language, Speech, and Mind*, New York: Routledge, pp. 3-29, 1988.

[23] Verdonck-de Leeuw, I. and Mahieu, H.F. "Vocal aging and the impact on daily life: A longitudinal study", *Journal of Voice*, Vol. 18, 2, pp. 193-202, 2004.

[24] Thomas, L.B., Harrison, A.L., and Stemple, J.C., "Aging thyroarytenoid and limb skeletal muscle: Lessons in constrast", *Journal of Voice*, V. 22, pp. 430-450, 2008.

[25] J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T. Toledano and J. Gonzalez-Rodriguez, "ATVS-UAM NIST SRE 2010 System", in *Proceedings of FALA 2010*, Vigo, Spain, November 2010.

[26] Kreiman, J. and Sidtis, D., *Foundations of Voice Studies*, New York: Wiley, 2011.

[27] Boersma, P. and Weenink, D., *Praat: Doing Phonetics by Computer.* [Computer Software], Amsterdam: Department of Language and Literature, University of Amsterdam, 2014.

[28] Colton, R.H. and Hollien, H., "Perceptual differentiation of the modal and falsetto registers", *Folia Phoniatrica et Logopaedica*, Vol. 25, n.4, pp. 270-280, 1973.

[29] Ingle, J. K., Wright, R. A., and Wassink, A. B., "Pacific Northwest vowels: A Seattle neighborhood dialect study", Paper presented at the *149th Meeting of the Acoustical Society of America*, Vancouver, 2005. http://www.aip.org/149th/ingle.html

[30] Künzel, H., Masthof, H.R. and Köster, J-P., "The relation between speech tempo, loudness, and fundamental frequency: An important issue in forensic speaker recognition", *Science and Justice*, Vol. 35, pp. 291-295, 1995.

[31] Künzel, H., "Some general phonetic and forensic aspects of speaking tempo", *Forensic Linguistics*, Vol. 4, n.1, pp. 48-84, 1997.

[32] Braun, A. and Schilz, J., "Speaking tempo in forensic phonetics", abstract presented at *IAFPA Conference 2009*, Cambridge, United Kingdom, 2009.

[33] Dellwo, V., Leemann, A., Kolly, M.-J., and Meyer, M., "Auditory speaker identification based on suprasegmental temporal characteristics". Abstract presented at IAFPA Conference 2013, Tampa, USA, 2013.

[34] Jessen, M., "Forensic reference data on articulation rate in German, *Science and Justice*, Vol. 47, pp. 50-67, 2007.

[35] Honglin Cao & Yingli Wang, "A forensic aspect of articulation rate variation in Chinese" *Proc. of the XVII ICPh.Sc.*, Hong-Kong, pp. 396-399.

[36] Van Dommelen, W.A., "The contribution of speech rhythm and pitch to speaker recognition", *Language and Speech*, Vol. 34, n. 4, pp. 325-338, 1997.

[37] Abberton, E. and Fourcin, A.J. Intonation and speaker identification", *Language and Speech*, Vol. 21, pp. 305-318, 1978.