Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 159234, 11 pages
doi:10.1155/2009/159234

Research Article

Analysis of Acoustic Features in Speakers with Cognitive Disorders and Speech Impairments

Oscar Saz,[1] Javier Simón,[2] W.-Ricardo Rodríguez,[1] Eduardo Lleida,[1] and Carlos Vaquero[1]

[1] *Communications Technology Group (GTC), Aragón Institute for Engineering Research (I3A), University of Zaragoza, 50018 Zaragoza, Spain*
[2] *Department of General and Hispanic Linguistics, University of Zaragoza, 50009 Zaragoza, Spain*

Correspondence should be addressed to Oscar Saz, oskarsaz@unizar.es

Received 31 October 2008; Revised 11 February 2009; Accepted 8 April 2009

Recommended by Juan I. Godino-Llorente

This work presents the results in the analysis of the acoustic features (formants and the three suprasegmental features: tone, intensity and duration) of the vowel production in a group of 14 young speakers suffering different kinds of speech impairments due to physical and cognitive disorders. A corpus with unimpaired children's speech is used to determine the reference values for these features in speakers without any kind of speech impairment within the same domain of the impaired speakers; this is 57 isolated words. The signal processing to extract the formant and pitch values is based on a Linear Prediction Coefficients (LPCs) analysis of the segments considered as vowels in a Hidden Markov Model (HMM) based Viterbi forced alignment. Intensity and duration are also based in the outcome of the automated segmentation. As main conclusion of the work, it is shown that intelligibility of the vowel production is lowered in impaired speakers even when the vowel is perceived as correct by human labelers. The decrease in intelligibility is due to a 30% of increase in confusability in the formants map, a reduction of 50% in the discriminative power in energy between stressed and unstressed vowels and to a 50% increase of the standard deviation in the length of the vowels. On the other hand, impaired speakers keep good control of tone in the production of stressed and unstressed vowels.

1. Introduction

The presence of certain speech and language disorders produces a decrease in the intelligibility of the speech in the patients affected with them [1]. In languages like Spanish, vowels are the nuclei of every syllable and play an important role in the intelligibility of speech, so the decrease in their quality and discriminative power has a major effect in the overall intelligibility of the speech. The goal of this work is to analyze and characterize this loss of intelligibility in a group of young speakers with cognitive disorders.

Several analytic studies have been carried out in studying the vocalic production of patients with different speech impairments. Cases of aphasia, disorder in the language due to brain damage, have been studied to understand their influence and the decrease of quality in the vocalic production [2, 3] by these patients. Dysarthria has been also studied where claims of patients with severe affections still controlling some of their suprasegmental vocalic features have been made [4], although with a lack of fine control over them. The affection to vocalic production in speech disorders due to Down's syndrome has been also studied [5] in pre- and postsurgical situations. Finally, the authors did an initial approach to this kind of analysis [6, 7] with the Spanish database of project HACRO containing different kinds of impaired speech [8].

In this work, it will be studied how vowel production quality varies in a group of young speakers with cognitive disorders and, sometimes severe, speech impairments associated to them like dysarthria, with respect to a set of reference unimpaired speakers. Four features will be studied: formant frequencies, fundamental frequency (tone), intensity (energy) and duration (length). Formants are the acoustic parameters required to distinguish different vowels, while tone and intensity may play the main role in the utterance of stressed versus unstressed vowels [9, 10]. Finally,

duration of vowels affects the correct perception of syllable prominence and position within the whole word or utterance [11], although its impact is not clear in Spanish language.

The organization of this paper is as follows: In Section 2, the acoustic features to be studied in this work will be presented from the point of view of acoustic and perceptual phonetics. Section 3 will introduce the young speech corpora used in this paper: the reference subcorpus and the impaired subcorpus. In Section 4 the methods for the extraction of all the studied features will be presented, as well as the reference values extracted from the unimpaired speech corpus. The results over the impaired speech corpus and the comparison with the reference values will be given in Section 5 and discussed in Section 6. Finally, the conclusions to this work will be extracted in Section 7.

2. Features of Spanish Vowels

This section will give a brief review on the main acoustic features of the vocalic production, focusing on their influence on the articulation of the Spanish vowels. The Spanish language contains five vowels (/*a*/, /*e*/, /*i*/, /*o*/ and /*u*/) clearly defined by their position in the formants map as it will be shown in the study of the reference corpus in Section 4.1. There are two allophones of the /*i*/ and /*u*/ vowels acting like glides ([*j*] and [*w*], resp.) that, despite being close to the vowels, cannot be considered as vocalic sounds when they are unstressed vowels and make the transition to a purely vocalic sound which is the nucleus in the syllable [12]. Hence, these glides are never considered for analysis in this work. Next, we will provide a basic theory of the Spanish vowels, according to their acoustic production and their influence in the perception of speech.

2.1. Formants. Formant frequencies are the only acoustic feature needed to describe Spanish vowels, where these frequencies rely heavily on the articulatory properties of each vowel [13]. The two main articulatory properties are the horizontal position of the tongue (defining palatal or front versus velar or back vowels) and the vertical position of the tongue (defining high versus low vowels). With this classification, a low position of the tongue will produce a higher first formant; while a more palatal position of the tongue will produce a higher second formant. Higher order formants like the third or fourth formants do not have a significant impact in Spanish vowels and are not considered in this work; moreover, tone doesn't have an impact either in the distinction of vowels.

According to this organization, Spanish has two high vowels (low first formant, 300–400 Hz): the velar /*u*/ (low second formant, 900 Hz) and the palatal /*i*/ (high second formant, 2300–2700 Hz), while only one low vowel (high first formant, 700–900 Hz) /*a*/ with a central position between palatal and velar (middle second formant, 1500–1700 Hz). Finally, two more vowels share a central-high position (high first formant, 500–600 Hz): the velar /*o*/ (low second formant, 1000–1200 Hz) and the palatal /*e*/ (high second formant, 2000–2400 Hz) [14].

2.2. Suprasegmental Features. There are three main acoustical features that affect the suprasegmental production in Spanish: tone, intensity and duration. In isolated words like it is the case of the work in this paper, these features mostly affect the distinct perception of stressed and unstressed vowels, although they do it in very different ways. Stress is considered in many phonetic theories as a binary feature that can be characterized as +stress or –stress, as perceived by the listener. Several trends differ in which suprasegmental feature carries most of the stress information, although nowadays it is widely accepted that tone is the main carrier of stress [15], followed by intensity. Anyways, no categorical assertion can be made in this subject, as the main prosody of the sentence and other microprosodic features can affect this perception in different utterances, as well as in the different characterization of tone in each language.

Finally, duration also has an influence in the perception of stress, but it is very affected by the fact that every syllable has a canonic length, so the duration of a stressed vowel is only comparable to the duration of the same unstressed vowel when they are the nucleus of the same syllabic structure. Otherwise, no categorical conclusion can be made from the comparison of the duration of stressed and unstressed vowels.

3. Corpora for Analysis

This section will present the most interesting features of the corpora used in this work for the analysis carried out in Sections 4 and 5. Further information concerning other features of the corpus can be found in [16]. The vocabulary used in the recording sessions is the 57 words from the Induced Phonological Register (RFI) [17], a very well-known speech therapy handbook in Spanish. These 57 words contain 129 syllables and 292 phonemes, with several repetitions of the vowels in different syllabic structures (90 different syllables). More precisely, the total number of vowels in the set of words is 129 (58 /*a*/, 18 /*e*/, 9 /*i*/, 38 /*o*/ and 6 /*u*/), each one of them being the nucleus of each one of the 129 syllables (in Section 2, it was argued how glides are considered nonvocalic sounds).

The process of the speech acquisition was made using "Vocaliza" [18]; this computer-aided speech therapy tool allows the acquisition of speech elicited from children prompting them with text, audio and images. Recordings were made in an empty classroom environment with a close-talk microphone (AKG C444L) connected to a laptop with a conventional sound card acquiring the signals in 16 kHz sampling frequency and storing them with a depth of 16 bits. The main corpus is divided into two subcorpora: unimpaired and impaired speech.

3.1. Unimpaired Speech Corpus. The unimpaired speech subcorpus contains speech from 168 young speakers (73 males and 95 females) in the range of 10 to 18 years old attending classes at primary and secondary school in Zaragoza, Spain. Every speaker has uttered one session of the isolated words in the RFI. The total number of utterances

in this subset of the corpus is 9576 isolated words (6 hours of signal). Recording process was fully supervised by at least a member of the research team to assure the good quality of the pronunciation and intelligibility of the utterances. Furthermore, only children with a good literacy assessment by their teachers were chosen to take part in the recordings. This subcorpus was recorded with the idea of providing a reference in the standard features in the speech of young speakers as it is well known that children' speech has special features [19].

3.2. Impaired Speech Corpus. The impaired speech subset of the corpus contains speech from 14 young speakers whose relation in terms of age and gender is shown on Table 1. Every speaker has uttered 4 sessions of the RFI isolated words; this is, 228 isolated words per speaker and a total of 3192 isolated words in the corpus (3 hours of signal). All 14 speakers suffer from cognitive disabilities and sometimes are also physically handicapped [16]. These disabilities affect their speech, producing a decrease in the quality and intelligibility in their utterances and also severe mispronunciations of some phonemes, which are either substituted by another phoneme or completely deleted.

Every utterance in the impaired speech subcorpus was manually labeled by three different experts to determine the perception of pronunciation mistakes made by the speakers. With a pairwise interlabeler agreement of 89.65% the mispronunciation rate (substitutions and deletions) is 17.61% for the overall set of phones (vowels, glides and consonants). The results in vowel mispronunciation per speaker are shown on Table 2, where it can be seen how there is a great variability in the affection to every speaker's speech, with some speakers making nearly no mistakes, while some others reaching 20% of mistakes. Although some speakers are not making any mistakes in the vowels, this does not indicate that their voice is completely healthy, because they present some degree of dysarthria that affects their voice quality.

Average mispronunciation rate of every vowel is shown in Table 3; the mean result for the 5 vowels altogether is 7.43% of mispronunciations, where /*a*/ and /*o*/ are around 4%-5% and /*e*/, /*i*/ and /*u*/ are more frequent mispronounced with 9-10% of mistakes. Once again, it is to remark that this manual labeling only refers to the substituted and deleted phonemes, resembling a perceptual labeling of how human experts perceive the phonemes (as the canonical one or as any other, but not indicating which was the actual phoneme uttered by the speakers in substitution of the canonical expansion).

Table 1: Impaired speakers in the corpus (Down's stands for Down's Syndrome).

Speaker	Age	Gender	Degree	Speaker	Age	Gender	Degree
Spk01	13	Female	Down's	Spk02	11	Male	Severe
Spk03	21	Male	Moderate	Spk04	20	Female	Moderate
Spk05	18	Male	Down's	Spk06	16	Male	Moderate
Spk07	18	Male	Severe	Spk08	19	Male	Severe
Spk09	11	Female	Moderate	Spk10	14	Female	Moderate
Spk11	19	Female	Moderate	Spk12	18	Male	Severe
Spk13	13	Female	Down's	Spk14	11	Female	Moderate

Table 2: Rate of vowel mispronunciations per speaker.

Speaker	Spk01	Spk02	Spk03	Spk04	Spk05	Spk06	Spk07
%Errors	0.39%	3.10%	0.39%	0.39%	17.44%	0.19%	0.78%
Speaker	Spk08	Spk09	Spk10	Spk11	Spk12	Spk13	Spk14
%Errors	8.53%	0.78%	7.56%	3.10%	8.33%	28.68%	0.00%

Table 3: Rate of mispronunciations per vowel.

Vowel	/*a*/	/*e*/	/*i*/	/*o*/	/*u*/
% Errors	4.16%	9.92%	9.52%	4.61%	8.93%

4. Acoustic Analysis and Reference Results

The acoustic analysis carried out aims to achieve a robust estimation of the four features concerned for study explained in Section 2. This Section gives a brief review of the algorithms used for the acoustic analysis and focuses on the reference results over the unimpaired subcorpus. State-of-the-art speech processing algorithms are implemented to estimate these values following the diagram on Figure 1 as also implemented in the speech therapy tool "PreLingua" for the improvement of phonatory controls in young children [20]. The speech processing is applied framewise (with a frame length of 25 milliseconds. and a frame shift of 10 milliseconds.) after obtaining the automated segmentation of the input speech via a Viterbi-based forced alignment. Hidden Markov Models (HMMs) used for the Viterbi alignment were trained with 3 different databases containing adult unimpaired speech: Albayzin [21], SpeechDat-Car [22] and Domolab [7]. 39-dimension Mel Frequency Cepstral Coefficients (MFCCs) vectors are used as features for the HMM alignment, composed of 12 static features and energy plus delta features plus delta-delta features. An example of the outcome of the automated segmentation over one of the utterances in the unimpaired children's subcorpus can be seen in Figure 2(a). The automated segmentation is initially based on the canonic transcription of every one of the utterances (isolated words) but, to avoid the pernicious effect of phoneme deletions in the impaired speakers' pronunciations, the deleted phonemes (as perceived in the human labeling) are not fed as input into the automated segmentation, as shown in the example in Figure 2(b).

After segmentation, impaired speech will be studied in two different groups: correctly pronounced vowels and mispronounced vowels. This way, the intelligibility will be studied separately in the situations in which the labelers still understand the vowel as correctly pronounced and in the situation of perception of mispronunciations.

4.1. Feature Estimation. The feature estimation is carried out following the next steps: after signal preprocessing (DC offset, pre-emphasis and Hamming windowing), a Linear Prediction Coefficient (LPC) analysis [23] is applied to every

Table 4: Mean ($\boldsymbol{\mu}$), standard desviation ($\boldsymbol{\sigma}$), skewness ($\boldsymbol{\gamma}_1$) and excess Kurtosis ($\boldsymbol{\gamma}_2$) values for the first and second formants in the reference corpus.

	First formant				Second formant			
Vowel	$\boldsymbol{\mu}$	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
/a/	762.75	108.77	−0.29	−2.90	1567.30	288.48	0.27	−0.78
/e/	512.21	61.73	−0.21	−3.00	2356.78	422.64	0.45	0.16
/i/	379.58	68.17	0.52	−2.98	2787.75	267.27	0.14	−0.16
/o/	552.72	69.46	−0.23	−2.95	1173.13	212.38	1.31	2.56
/u/	423.40	61.48	−0.26	−2.98	1083.16	213.67	0.69	0.32

Table 5: Mean ($\boldsymbol{\mu}$), standard desviation ($\boldsymbol{\sigma}$), skewness ($\boldsymbol{\gamma}_1$) and excess Kurtosis ($\boldsymbol{\gamma}_2$) values of pitch in the reference corpus (Females 13-14 years-old).

	Stressed vowels				Unstressed vowels			
Vowel	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
/a/	229.67	30.43	−0.54	1.04	207.78	27.54	0.65	4.02
/e/	229.04	30.12	−0.39	0.89	219.37	29.08	0.02	0.42
/i/	241.08	33.38	−0.54	1.30	219.51	29.64	0.74	3.71
/o/	228.55	32.45	−0.53	0.93	203.24	26.42	0.41	2.65
/u/	237.66	39.23	−0.44	1.30	236.88	30.68	−0.28	2.02

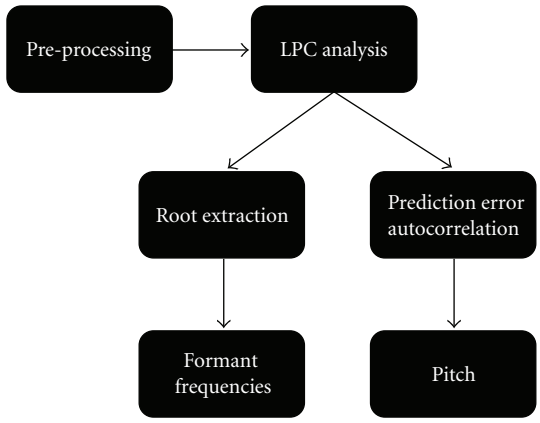


Figure 1: Acoustic analysis diagram

frame to extract the roots of coefficients (a_k) of the 16-order speech prediction model in (1).

$$H(z) = \frac{G}{1 - \sum_{k=1}^{16}(a_k z^{-k})}, \tag{1}$$

where the input signal $s(n)$ is estimated as $\hat{s}(n)$ using the time-domain impulsional response $h(n)$ associated to H(z) as in (2):

$$\hat{s}(n) = h(n) * s(n). \tag{2}$$

The estimation of the formants takes the 16 LPC coefficients (a_k) in the prediction model $H(z)$ and extracts the polynomial roots, each one of them associated to a formant frequency. The roots with the two higher absolute values will correspond to the first and second formants.

Tone estimation calculates the autocorrelation of the prediction error $e(n)$ given in (3) and its autocorrelation $r(k)$ in (4) with fr_l the value of frame length (25 milliseconds per frame):

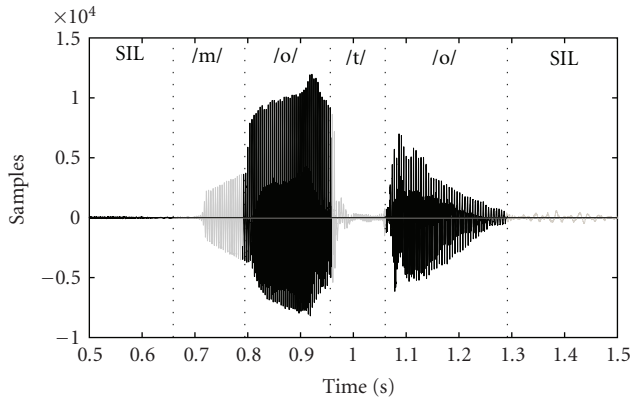
$$e(n) = s(n) - \hat{s}(n), \tag{3}$$

$$r(k) = \sum_{n=0}^{fr\_l} e(n)e(n-k), \tag{4}$$

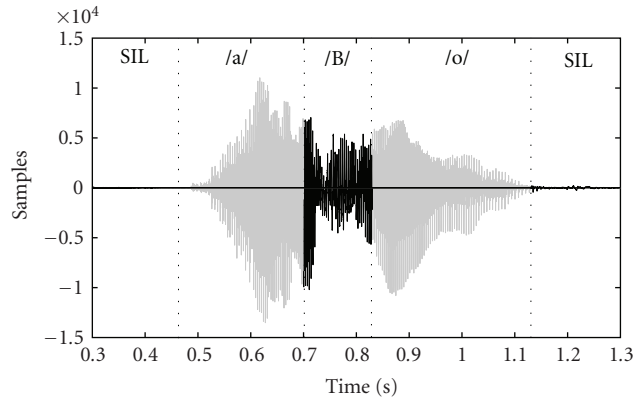
The index k in which the autocorrelation has its maximum value outside from the area around the origin $r(0)$ will be the pitch period (k_{pitch}) associated to the pitch frequency ($F_{\text{pitch}} = F_{\text{sample}}/k_{\text{pitch}}$) where F_{sample} is 16 kHz as mentioned before. An estimation of the sonority value, as the ratio between the maximum value of autocorrelation and the autocorrelation in the origin ($r(k_{\text{pitch}})/r(0)$), will indicate if the frame is sonorant enough to be considered as a vowel and, hence, use the calculated pitch and formant values as correct. A high sonority ratio avoids the possibility of pitch and formant prediction mistakes, although some correct frames might be rejected.

For the intensity estimation, some arguments have to be considered. First, actual values of intensity (this is, sample values or directly computed frame energy) cannot be considered into the study as it is not possible to reliably argue that input intensity during the recording process stayed steady through all different sessions, as the recordings of all the speakers took more than a year. However, it is reasonable to argue that Signal-to-Noise Ratio (SNR) will maintain constant for similar speech intensity independently of the input volume since a close-talk microphone was used for the recordings.

This assumption is evaluated by the estimation of the background noise power level calculated for the corpus used in the work, whose mean value is 27.15 dB (7.22 dB of

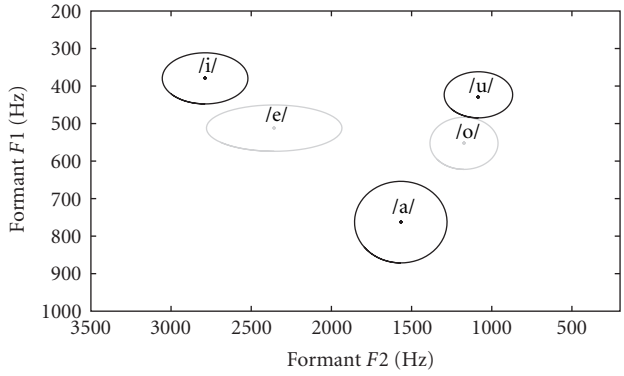


(a) Utterance of the word "moto" (SIL-/*m*/-/*o*/-/*t*/-/*o*/-SIL) by an unimpaired male of 11 years old.

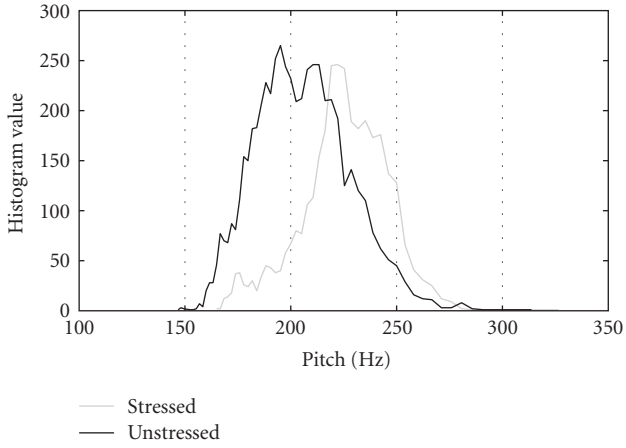


(b) Utterance of the word "árbol" (SIL-/*a*/-/*r*/-/*B*/-/*o*/-/*l*/) by Spk05 where /*r*/ and /*l*/ are labeled as deletions.

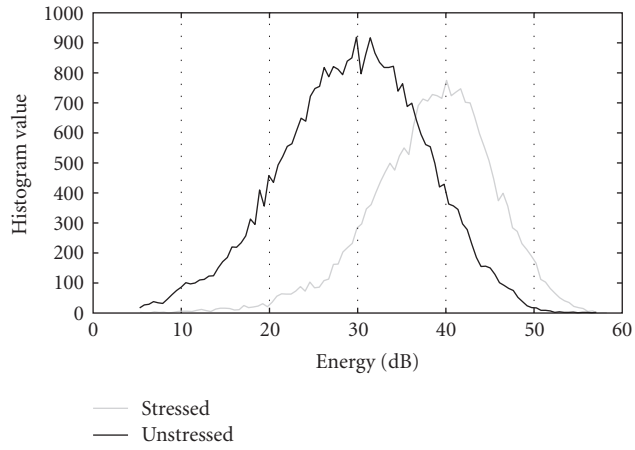
Figure 2: Examples of the outcome of the automated segmentation



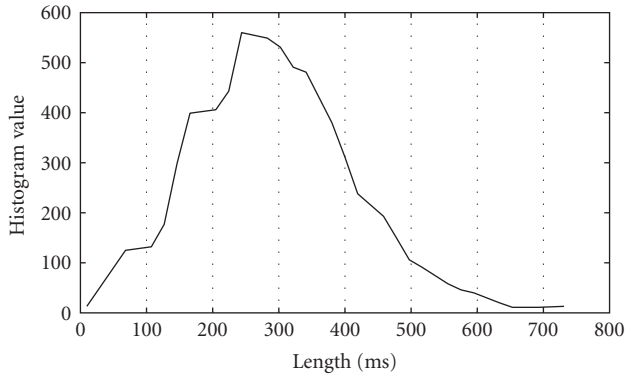
(a) Formant representation (mean and standard deviation) in the reference corpus.



(b) Pitch histogram for stressed and unstressed vowels /*o*/ in the reference corpus (females 13-14 years old).

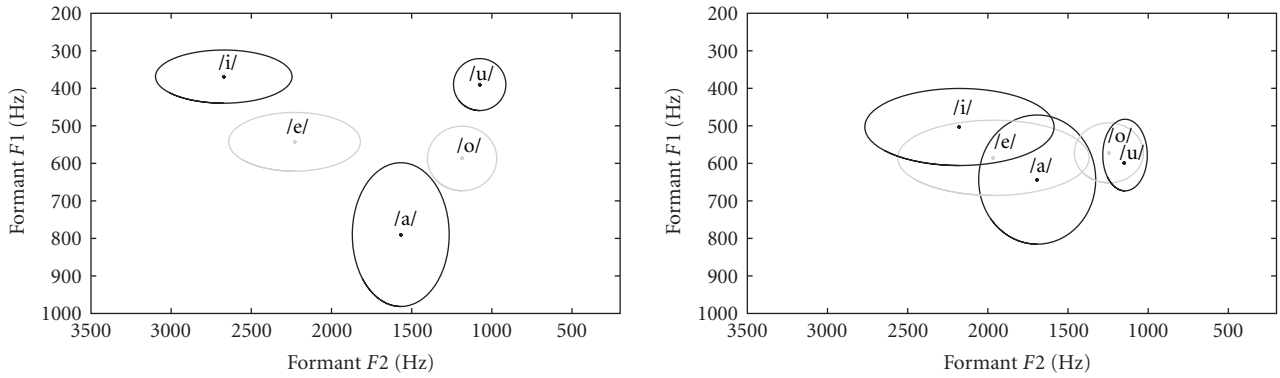


(c) Energy histogram of stressed and unstressed vowels /*o*/ in the reference corpus.



(d) Length histogram of vowels /*o*/ in the reference corpus.

Figure 3: Representation of the 4 features: (a) Formants, (b) Pitch, (c) Energy, (d) Length



(a) Formant representation (mean and standard deviation) in the impaired corpus (Correct vowels)

(b) Formant representation (mean and standard deviation) in the impaired corpus (Mispronounced vowels)

Figure 4: Formants map for the impaired speakers

Table 6: Mean ($\boldsymbol{\mu}$), standard desviation ($\boldsymbol{\sigma}$), skewness ($\boldsymbol{\gamma}_1$) and excess Kurtosis ($\boldsymbol{\gamma}_2$) values of the frame wise energy (SNR) in the reference corpus.

	Stressed vowels				Unstressed vowels			
Vowel	μ	σ	γ_1	γ_2	μ	σ	γ_1	γ_2
/a/	37.78	6.93	−0.39	0.39	30.34	8.32	−0.27	−0.16
/e/	37.21	7.36	−0.58	0.97	34.18	7.91	−0.27	0.15
/i/	36.77	6.44	−0.33	−50	33.18	7.29	−0.39	0.44
/o/	38.42	6.96	−0.57	0.78	29.46	8.11	−0.18	−0.17
/u/	37.27	7.12	−0.46	0.34	34.61	6.34	−0.35	0.30

Table 7: Mean ($\boldsymbol{\mu}$), standard desviation ($\boldsymbol{\sigma}$), skewness ($\boldsymbol{\gamma}_1$) and excess Kurtosis ($\boldsymbol{\gamma}_2$) values of the vowel length in the reference corpus.

Vowel	μ	σ	γ_1	γ_2
/a/	120.75	53.11	1.25	4.49
/e/	107.73	50.81	1.06	2.45
/i/	114.88	39.42	0.56	0.62
/o/	123.03	58.42	2.24	13.25
/u/	113.16	45.89	0.62	0.52

standard deviation) for the reference subcorpus and 27.07 dB (6.61 dB of standard deviation) for the impaired subcorpus, which validates the hypothesis that noise level is directly related to intensity level and maintains similar and good properties through all the recordings. Hence, prior to energy estimation, average background noise power is calculated through all the frames considered as nonspeech in the forced alignment. Afterwards, for each frame of the vowels, framewise energy is calculated and SNR is obtained by subtracting the noise power in the utterance. For convention purposes, from now on, intensity or energy will be this value of SNR where the background noise level has been substracted.

Duration calculation is done by estimating the length of the vowel in milliseconds, computing the number of frames assigned to each vowel in the forced alignment and then multiplying by the frame shift value of 10 milliseconds per frame. A threshold over the energy is applied to restrict the vowel boundaries and hence avoid the effect of coarticulation in the transitions to or from consonantal sounds. This threshold was preset to restrict boundary frames with low energy whose calculation of pitch and formants could be inaccurate.

4.2. Reference Results. The reference subcorpus of 168 unimpaired young speakers was initially analyzed to determine the standard values of the formants and suprasegmental features under study in this work. Some general assumptions will be made in this paper concerning the statistical properties of the features studied in this work: First, the values of the formants have a 2-dimension Gaussian distribution for each vowel. Values of pitch and energy have a Gaussian distribution separately for stressed and unstressed vowels (pitch can only be considered for one speaker alone or for a population of the same gender and age). Finally, the values of vowel length have a Gaussian distribution for each vowel.

All the values in this Section are given in terms of mean (μ), standard deviation (σ), skewness (γ_1) and excess Kurtosis (γ_2); where the values (close to zero) of γ_1 and γ_2 validate the Gaussian assumptions. Once assured the Gaussian properties, in the studies on the impaired subcorpus in Section 5, μ and σ will be the only statistics. All reference

Table 8: Mean (μ) and standard desviation (σ) for the formants in the impaired corpus.

	Correct vowels				Mispronounced vowels			
	First formant		Second formant		First formant		Second formant	
Vowel	μ	σ	μ	σ	μ	σ	μ	σ
/a/	789.90	191.70	1568.03	302.28	643.25	172.22	1692.28	365.08
/e/	542.55	78.29	2230.34	410.96	585.28	100.29	1965.08	597.43
/i/	368.81	70.89	2671.75	426.03	503.10	102.72	2177.16	590.60
/o/	586.92	86.15	1185.30	216.18	572.12	79.85	1245.47	215.34
/u/	390.02	69.24	1075.79	163.38	577.96	95.51	1144.39	130.56

Table 9: Mean (μ) values of pitch in the impaired corpus for correct vowels.

	Group A		Group B		Group C		Group D	
Vowel	Stressed	Unstressed	Stressed	Unstressed	Stressed	Unstressed	Stressed	Unstressed
/a/	150.93	139.10	231.46	200.71	259.56	233.79	308.50	278.65
/e/	152.43	145.35	238.40	208.26	259.42	248.60	302.11	283.32
/i/	170.45	144.40	254.82	219.44	277.78	242.95	308.34	283.63
/o/	150.94	139.27	236.52	200.28	259.28	232.22	302.90	270.11
/u/	161.93	146.68	251.93	230.36	267.58	245.53	316.96	292.26

values are shown on Tables 4 (formants), 5 (pitch), 6 (energy-SNR) and 7 (length). Table 5 shows only the results for the group of unimpaired females of 13-14 years old as an example of pitch trend in the unimpaired data (rest of groups behave similarly and are not shown here to restrict space of this Section, it is to remember that pitch has to be studied separately for gender and age to maintain the condition of Gaussian distribution).

A graphical representation of these features is given in Figure 3. First, Figure 3(a) shows the results of the formant analysis done over the reference corpus plotting and ellipsoid whose center are the mean values for the first and second formant and the axes are the standard deviations of both formants. Figures 3(b) and 3(c) show the histograms in the pitch and energy respectively for the vowel /o/ in the reference corpus, separating stressed vowels from unstressed vowels; while Figure 3(d) shows the histogram of the duration of the vowel /o/ across the reference corpus. Vowel /o/ has been chosen to provide this graphical view of the histograms for being one of the vowels with more appearances in the corpus.

Referring to the formant results in Table 4 and Figure 3(a), the values are similar to the canonic formant values accepted traditionally in Spanish phonetics, and a good discrimination can be made among all five vowels. Pitch and energy in Table 5 and 6 and Figures 3(b) and 3(c) show their discriminative effect in the perception of stress, as the pitch in stressed vowels is 10–20 Hz over the pitch of unstressed vowels and the energy in stressed vowels is 4-7 dB over the energy of unstressed vowels. Finally, regarding length in Table 7 and Figure 3(d), it is seen as vowel production is steady in its length, with a standard deviation not exceeding the 40%–50% of the mean results in length (around 120 milliseconds).

5. Impaired Speech Results

In this section, the results achieved in the acoustic analysis over the impaired speech subset of the corpus will be given. This analysis will comprehend the four acoustic features considered in Section 2, while making an initial comparison with the results in Section 4.2 over the reference subcorpus. The full comparative analysis will be made in Section 6 with the help of statistical tools like the Kullback-Leibler Divergence and the Fisher Ratio.

5.1. Formant Results. The formant map for the 14 impaired speakers is shown on Figure 4. Figure 4(a) provides the formant map for the vowels perceived as correctly pronounced by the human labelers, with their statistics given in the first columns of Table 8. Two major effects can be appreciated: First, the increase in the area of every vowel in the formant map in Figure 4(a), which is appreciated as an increase in the standard deviation of the formants in Table 8 when compared to the formants of the reference speakers in Table 4. And second, the approximation of vowels /a/, /e/ and /o/ towards the center of the formants map in Figure 4(a), also appreciated in the mean results in Table 8.

Concerning the results for the vowels perceived as mispronounced by the human labelers, given in Figure 4(b) and the second half of Table 8, there can be appreciated the total confusion in the formants, as expected in this case where a mistake in the pronounced vowel has been made by the speakers. In this case, all the formants are centered in the middle of the formant map and the standard deviation is much higher. In this case, what the speakers are really uttering is different from the canonical vowel to be expected and the production of speech is blurred in the formant map, as the labelers were not told to indicate what the speaker was really saying.

Table 10: Mean (μ) values of pitch in the impaired corpus for mispronounced vowels.

	Group A		Group B		Group C		Group D	
Vowel	Stressed	Unstressed	Stressed	Unstressed	Stressed	Unstressed	Stressed	Unstressed
/a/	150.27	137.56	263.48	229.33	267.54	256.01	264.44	236.27
/e/	151.98	146.10	246.23	228.11	255.77	268.71	306.03	312.17
/i/	139.74	149.60	272.98	210.56	270.37	225.41	—	—
/o/	150.75	138.28	241.19	210.61	260.71	245.16	—	250.61
/u/	—	138.04	223.33	196.08	273.40	250.00	—	267.78

Table 11: Mean (μ) and standard desviation (σ) of energy in the impaired corpus.

	Correct vowels				Mispronounced vowels			
	Stressed		Unstressed		Stressed		Unstressed	
Vowel	μ	σ	μ	σ	μ	σ	μ	σ
/a/	37.09	7.94	31.87	9.06	35.85	7.62	35.03	9.70
/e/	37.93	7.29	34.02	8.34	35.25	8.20	32.88	7.37
/i/	37.59	7.85	33.36	7.60	37.37	9.55	30.98	8.82
/o/	37.91	8.47	34.70	9.69	37.01	9.58	33.22	8.69
/u/	38.63	7.81	34.38	7.03	36.86	9.64	27.84	8.50

Table 12: Mean (μ) and standard desviation (σ) of length in the impaired corpus.

	Correct vowels		Mispronounced vowels	
Vowel	μ	σ	μ	σ
/a/	138.42	75.62	99.47	109.53
/e/	142.01	84.72	100.17	87.69
/i/	128.88	66.76	143.75	117.11
/o/	151.66	93.81	115.33	120.13
/u/	127.42	64.62	138.40	77.98

5.2. Tone (Pitch) Results. The study of the pitch values for the impaired subcorpus should best given separately for every speaker; however, the lack of sufficient data for a correct statistical analysis (especially when studying mispronounced vowels) leads to the need of gathering speakers in groups with similar pitch values. Hence, 4 groups are created,

(i) Group A gathers speakers Spk03, Spk06, Spk07 and Spk12 (4 of the older males with very low pitch values).

(ii) Group B gathers speakers Spk05, Spk08, Spk10 and Spk11 (2 females and 2 males with a medium pitch values).

(iii) Group C gathers speakers Spk04, Spk09, Spk13 and Spk14 (4 females with a medium-high pitch values)

(iv) Group D gathers speakers Spk01 and Spk02 (male and female with a high pitch).

The results for the 4 groups of speakers are given in Tables 9 (correctly pronounced vowels) and 10 (mispronounced vowels, where some values are missing due to the not existence of data for those cases).

It can be seen as impaired speakers keep a good control of these prosodic features: Values of pitch are steady among all five vowels and speakers show the ability to discriminate stressed vowels from unstressed vowels in all 5 vowels in similar ways to reference speakers (with 10–20 Hz of separation between stressed and unstressed vowels). We have to consider with caution the results in the case of mispronounced vowels, as the nonexistence of some cases leads to strange results.

5.3. Intensity (Energy) Results. Regarding the values of framewise energy (SNR as explained on Section 4), the average results for all the impaired speakers are given in Table 11. It is seen how energy keeps good properties for the impaired speakers, and they are able to produce an increase in their intensity production when uttering stressed vowels, although compared to the reference results in Section 4.2 there is a slight increase in the energy of unstressed vowels. On the other hand, a reduction in the energy in stressed vowels is noticed in the vowels labeled as mispronunciations.

5.4. Duration (Length) Results. The statistics for the results of the vowel length in the group of 14 impaired speakers are shown on Table 12. It can be seen that there is an increase in the average length of around 15 milliseconds for all vowels when compared to the reference speakers in Table 7, but what it is more noticeable is the increase in standard deviation (more than 50%), which indicates the presence of vowels with a very variable length, meaning the existence of extremely long and extremely short vowels, as there is no significant change in the skewness and Kurtosis of the statistics. The increase in standard deviation is especially noticeable in the mispronounced vowels, which indicates that what the speakers are really uttering instead of the vowels is a non steady realization of speech. This clearly might be

Table 13: sKLD and FR for the most confusable pairs of vowels in the Spanish formants (Avg is the weighted average over the number of appearances of every vowel).

Unimpaired speakers		Impaired speakers (Correct vowels)		Impaired speakers (Mispronounced vowels)		
Vowels	sKLD	FR	sKLD	FR	sKLD	FR
/a/-/e/	17.40	6.39	11.80	3.11	1.78	0.24
/a/-/o/	9.72	3.86	7.51	1.99	5.43	1.25
/e/-/i/	6.49	2.82	6.60	3.26	0.78	0.39
/o/-/u/	4.15	2.03	7.27	3.34	1.02	0.16
/e/-/o/	20.97	6.45	16.17	5.21	9.35	1.29
Avg.	12.67	4.62	10.10	3.19	4.17	0.76

Table 14: sKLD and FR in pitch between stressed and unstressed vowels (Avg is the weighted average over the number of appearances of every vowel).

Unimpaired speakers		Impaired speakers (Correct vowels)		Impaired speakers (Mispronounced vowels)		
Vowels	sKLD	FR	sKLD	FR	sKLD	FR
/a/	1.15	0.51	1.02	0.46	4.14	1.66
/e/	0.52	0.13	0.52	0.23	1.29	0.21
/i/	1.19	0.54	1.86	0.55	48.97	1.16
/o/	1.43	0.67	0.94	0.42	2.37	0.53
/u/	0.19	0.06	0.47	0.28	3.06	0.24
Avg	1.10	0.49	0.96	0.41	6.30	1.02

indicating that speakers are unsure of their production of speech, so they are trying to skip that vowel (making it shorter) or making it longer while they try to pronounce the right sound.

6. Discussion

The results obtained in Section 5 can give way to a discussion on several aspects of the vocalic production of impaired speakers. The discussion in this section will come accompanied with the computation of the Kullback-Leibler Divergence (KLD) and the Fisher Ratio (FR) [24]. These two measures are known to provide a good metric of the discriminative power of two different random variables. In this work, they will help to know the discriminative separation between vowels in the formant map and between stressed and unstressed vowels in terms of tone and intensity.

For this work, it will be considered the KLD definition for n-dimensional Gaussians distributions (2-dimensional in the case of formants and 1-dimensional in the other features). This definition, considered for two distributions $A \sim \aleph(\mu_A, \Sigma_A)$ and $B \sim \aleph(\mu_B, \Sigma_B)$ where μ_A and μ_B are mean vectors, Σ_A and Σ_B diagonal covariance matrices and n the dimension of the distributions, is given by (5):

$$KL(A,B) = \sum_{i=0}^{n} \left(\log\left(\frac{\Sigma_{A_i}}{\Sigma_{B_i}} \right) + \frac{\left(\mu_{A_i} - \mu_{B_i}\right)^2}{\Sigma_{B_i}} + \frac{\Sigma_{A_i}}{\Sigma_{B_i}} - 1 \right). \tag{5}$$

However, given this definition, the KLD is nonsymmetric, this means that $KL(A,B) \neq KL(B,A)$, so a symmetrized KLD (sKLD) is defined in (6):

$$sKLD(A,B) = \frac{KL(A,B) + KL(B,A)}{2}. \tag{6}$$

Finally, the FR equation for the two n-dimensional Gaussian distributions $A \sim \aleph(\mu_A, \Sigma_A)$ and $B \sim \aleph(\mu_B, \Sigma_B)$ is given in (7):

$$FR(A,B) = \sum_{i=0}^{n} \left(\frac{\left(\mu_{A_i} - \mu_{B_i}\right)^2}{\Sigma_{A_i} + \Sigma_{B_i}} \right). \tag{7}$$

Concerning the formants (the only acoustic feature of the vowels) there is an important decrease in sKLD and FR in the formant map between the vowels /a/, /e/ and /o/ in Table 13, while vowels /i/ and /u/ separate from the other 3 vowels, increasing their sKLD and FR in the formant map.

However, this is not a precise vision of the situation, because it is not to be forgotten than these two vowels are the less likely seen in Spanish language; not only in the vocabulary of this work in Section 3, but also in some other major text corpora in Spanish like the Europarl corpus [25], where the percentage of appearances of vowels is 11.83% for /e/, 9.51% for /a/, 8.07% for /o/ and only 4.28% for /i/ and 1.74% for /u/. This way, when computing a weighted average result in sKLD and FR (last row of Table 13), where the weights are the percentage of appearances of every vowel in the vocabulary, it is seen that there is an average reduction of 20.28% in the sKLD and 30.95% in the FR between

Table 15: sKLD and FR in energy between stressed and unstressed vowels (Avg is the weighted average over the number of appearances of every vowel).

Unimpaired speakers		Impaired speakers (Correct vowels)		Impaired speakers (Mispronounced vowels)		
Vowel	sKLD	FR	sKLD	FR	sKLD	FR
/*a*/	1.04	0.47	0.42	0.19	0.13	0.00
/*e*/	0.17	0.08	0.29	0.12	0.12	0.05
/*i*/	0.31	0.14	0.30	0.15	0.24	0.11
/*o*/	1.49	0.70	0.51	0.23	0.19	0.09
/*u*/	0.19	0.08	0.35	0.16	1.04	0.49
Avg	0.96	0.44	0.42	0.19	0.20	0.06

unimpaired and impaired speakers uttering correctly the vowels. This reduction in discriminative power rises up to 83.55% between unimpaired and impaired speakers in the situation of mispronunciations. This result is clearly expectable since we are considering a situation where the canonical form of the vowel has been not uttered, but it serves as a way of validating consistently the human labeling made by the experts.

In terms of suprasegmental features, the separation between stressed and unstressed vowels are given in Tables 14 (pitch) and 15 (energy). Table 14 shows that there is not a significant decrease in the weighted sKLD and FR in pitch between unimpaired speakers and impaired speakers (when uttering correctly the vowels). This corroborates previous works [4] in the fact that impaired speakers can still control some prosodic features in their speech even they lose intelligibility in their vowel production. The results in the mispronounced vowels by impaired speakers cannot be considered due to the pernicious effect of unseen cases in the test data.

It is in terms of energy (or intensity) where impaired speakers seem to have bigger problems in the control of prosody and stress. There is a reduction of 56.26% in sKLD and 56.82% in FR in the discriminative power between these two distributions, and this reduction increases to 80% in the case of mispronounced vowels. As mentioned in Section 5, this reduction in discriminative power is mostly due to an increase in the energy of unstressed vowels. The reason for that might be in the fact that impaired speakers are trying to assure themselves in their pronunciation by raising their intensity in their situations of hesitation. This extra intensity would not affect stressed vowels because stressed vowels have higher energy due to this prosodic feature of stress affecting them.

Finally, the study of the length of the production of vowels by the impaired speakers in Table 12 shows an effect of dispersion in the length of the vowels. This means that vowels as uttered by these speakers are more often abnormally long or short. Actually, two separate effects can be appreciated; in the case of correctly pronounced vowels by the impaired speakers there is an effect of lengthening of the vowels (around 20%–30% increase in mean values between Tables 7 and 12), while mispronounced vowels are excessively dispersed (with standard deviations of 80% of the mean values), mainly due to the doubts and hesitations of incorrect pronunciations. The increase of duration of correctly pronounced vowels might indicate certain hesitations in the speakers when uttering their speech, due to their insecurity in speech production because of their speech disorders.

7. Conclusion

As conclusions to this work, a whole corpus with unimpaired and impaired children' speech corpus has undergone an acoustic study based on LPC analysis to calculate acoustic features like formants and suprasegmental features (pitch, energy and length). Results show that the good properties of unimpaired speakers (well-behaved formants, separation of stressed and unstressed vowels in terms of pitch and energy, and statistically correct length features) are distorted in different ways in the impaired speakers.

Impaired speakers reduce in 20%–30% the discriminative ability of the formant map, even when the pronunciation is perceived as correct by a set of human experts. Results in the case of mispronunciations show a total blur in the formant map as expected and as detected by the human experts. Impaired speakers have a good control of tone as feature for the microprosody of the words; but intensity discrimination between stressed and unstressed vowels is reduced by a 50% due to an increase in the energy of unstressed vowels. Finally, it has been shown how these speakers have problems to maintain a steady production of vowels in terms of their length, with the abnormal production of extremely long or short vowels that is reflected in an increase of 50% in the standard deviation of the vowel length.

Hence, it can be concluded that the main problem in the vowel production due to the speech disorders analyzed in this work reflects in terms of formants, intensity control and vowel length, while they are able to maintain a correct production of pitch. Further work in this area may include a more precise analysis of the formant values, considering their relationship to the pitch value of every speaker. Also, the results in this work could be validated with the results achieved with a manual segmentation of vowels; although the automated segmentation is robust enough and, altogether

with the strict sonority threshold applied, assures that all the frames analyzed belong to vowels.

Further studies in the vowel duration may also be done considering a new vocabulary with the same syllables in different positions and situations of stress. Finally, a bigger study considering connected speech might be done to study the loss of prosody features in a situation of complete sentences. This study might be useful to determine if impaired speakers have problems with prosody control in a more complex context than simple control of stress features. Another study of interest would be to link these results to the outcome of a whole phonetic transcription of the speakers' speech (with a confusion matrix of the mispronunciations) and also analyze separately each speaker speech in terms of acoustic parameters, although that would require a more careful statistical analysis due to the reduction in the amount of data studied.

Acknowledgments

The authors want to acknowledge José Manuel Marcos, César Canalís, Pedro Pegero, and Beatriz Martínez from the School for Special Education "Alborada", located in Zaragoza (Spain), for their collaboration in this work.

References

[1] M.-J. Ball, *Phonetics for Speech Pathology*, Whurr Publishers, London, UK, 1983.

[2] K. Croot, "An acoustic analysis of vowel production across tasks in a case of non-fluent progressive aphasia," in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP-Interspeech '98)*, pp. 907–910, Sydney, Australia, December 1998.

[3] T. Prizl-Jakovac, "Vowel production in aphasia," in *Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech '99)*, pp. 583–586, Budapest, Hungary, September 1999.

[4] R. Patel, "Phonatory control in adults with cerebral palsy and severe dysarthria," *Augmentative and Alternative Communication*, vol. 18, no. 1, pp. 2–10, 2002.

[5] C. P. Moura, D. Andrade, L. M. Cunha, et al., "Voice quality in Down syndrome children treated with rapid maxillary expansion," in *Proceedings of the 9th European Conference on Speech Communication and Technology (EuroSpeech '05)*, pp. 1073–1076, Lisboa, Portugal, September 2005.

[6] O. Saz, A. Miguel, E. Lleida, A. Ortega, and L. Buera, "Study of time and frequency variability in pathological speech and error reduction methods for automatic speech recognition," in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP '06)*, pp. 993–996, Pittsburgh, Pa, USA, September 2006.

[7] R. Justo, O. Saz, V. Guijarrubia, A. Miguel, M.-I. Torres, and E. Lleida, "Improving dialogue systems in a home automation environment," in *Proceedings of the 1st International Conference on Ambient Media and Systems (Ambi-Sys '08)*, Québec, Canada, February 2008.

[8] J.-L. Navarro-Mesa, P. Quintana-Morales, I. Pérez-Castellano, and J. Espinosa-Yáñez, "Oral corpus of the project HACRO (Help tool for the confidence of oral utterances)," Tech. Rep., Department of Signal and Communications, University of Las Palmas de Gran Canaria, Gran Canaria, Spain, May 2005.

[9] O. Fujimura and D. Erickson, "Acoustic phonetics," in *The Handbook of Phonetic Sciences*, W.-J. Hardcastle and J. Laver, Eds., pp. 65–115, Blackwell, Oxford, UK, 1997.

[10] K.-N. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, Mass, USA, 1998.

[11] O. Scharenborg, "Modelling fine-phonetic detail in a computational model of word recognition," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '08)*, pp. 1473–1476, Brisbane, Australia, September 2008.

[12] J.-I. Hualde, *The Sounds of Spanish*, Cambridge University Press, Cambridge, UK, 2005.

[13] A. Quilis, *Fonética Acústica de la Lengua Española*, Gredos, Madrid, Spain, 1981.

[14] E. Martínez-Celdrán and A.-M. Fernández-Planas, *Manual de Fonética Española. Articulaciones y Sonidos del Español*, Ariel, Barcelona, Spain, 2007.

[15] D. Fry, "Experiments in the perception of stress," *Language and Speech*, vol. 1, pp. 126–152, 1958.

[16] O. Saz, W.-R. Rodríguez, E. Lleida, and C. Vaquero, "A novel children's corpus of disordered speech," in *Proceesing of the 1st Workshop on Child, Computer and Interaction (WOCCI '08)*, Chania, Greece, October 2008.

[17] M. Monfort and A. Juárez-Sánchez, *Registro Fonológico Inducido (Tarjetas Grficas)*, Cepe, Madrid, Spain, 1989.

[18] C. Vaquero, O. Saz, E. Lleida, and W.-R. Rodríguez, "E-inclusion technologies for the speech handicapped," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '08)*, pp. 4509–4512, Las Vegas, Nev, USA, March-April 2008.

[19] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[20] W.-R. Rodríguez, C. Vaquero, O. Saz, and E. Lleida, "Speech technology applied to children with speech disorders," in *Proceedings of the 4th International Conference on Biomedical Engineering (BioMed '08)*, pp. 247–250, Kuala Lumpur, Malaysia, June 2008.

[21] A. Moreno, D. Poch, A. Bonafonte, et al., "Albayzin speech database: design of the phonetic corpus," in *Proceedings of the 3th European Conference on Speech Communication and Technology (EuroSpeech '93)*, pp. 175–178, Berlin, Germany, September 1993.

[22] A. Moreno, B. Lindberg, C. Draxler, et al., "SpeechDat-Car: a large speech database for automotive environments," in *Proceedings of the 2nd Language Resources European Conference (LREC '00)*, Athens, Greece, June 2000.

[23] L.-R. Rabiner and R.-W. Schafer, *Digital Processing of Speech Signals*, Signal Processing Series, Prentice-Hall, Englewood Cliffs, NJ, USA, 1978.

[24] T.-M. Cover and J.-A. Thomas, *Elements on Information Theory*, Wiley Interscience, New York, NY, USA, 1991.

[25] P. Koehn, "Europarl: a parallel corpus for statistical machine translation," in *Proceedings of the 10th Machine Translation Summit*, pp. 79–86, Phuket,Thailand, September 2005.