Hypernasal speech detection by acoustic analysis of unvoiced plosive consonants

Alexander Sepúlveda Sepúlveda[1]

Edilson Delgado Trejos[2]

Santiago Murillo Rendón[3]

Germán Castellanos Domínguez[4]

Abstract

People with a defective velopharyngeal mechanism speak with abnormal nasal resonance (hypernasal speech). Voice analysis methods for hypernasality detection commonly use vowels and nasalized vowels. However to obtain a more general assessment of this abnormality it is necessary to analyze stops and fricatives.

This study describes a method with high generalization capability for hypernasality detection analyzing unvoiced Spanish stop

1 Ingeniero Electrónico. M. Sc en Automatización Industrial. Estudiante de doctorado en ingeniería LI Automática. Universidad Nacional de Colombia Sede Manizales. Email: fasepulvedas@unal.edu.co

2 Ingeniero Electrónico. M. Sc. en Automatización Industrial. Ph. D. en ingeniería LI Automática. Académico Investigador del Centro de Investigación, Instituto Tecnológico Metropolitano.
Email: edelgadot@unal.edu.co

3 Estudiante Ingeniería Electrónica. Universidad Nacional de Colombia Sede Manizales.
Email: smurillor@unal.edu.co

4 Ingeniero en Telecomunicaciones. Ph. D. En Ingeniería. Profesor asociado al Departamento de Ingeniería Eléctrica, Electrónica y Computación de la Universidad Nacional de Colombia Sede Manizales.
Email: cgcastellanosd@unal.edu.co

Fecha de recepción: 25 de septiembre de 2009
Fecha de aceptación: 13 de noviembre de 2009

consonants. The importance of phoneme-by-phoneme analysis is shown, in contrast with whole word parametrization which includes irrelevant segments from the classification point of view. Parameters that correlate the imprints of Velopharyngeal Incompetence (VPI) over voiceless stop consonants were used in the feature estimation stage. Classification was carried out using a Support Vector Machine (SVM), including the Rademacher complexity model with the aim of increasing the generalization capability. Performances of 95.2% and 92.7% were obtained in the processing and verification stages for a repeated cross-validation classifier evaluation.

Index Terms

acoustic analysis, speech analysis, hypernasality, unvoiced stop consonants and rademacher complexity.

Resumen

Las personas con un mecanismo velofaringeo defectuoso hablan con una resonancia nasal anormal (habla hipernasal). Métodos de análisis de voz para detección de hipernasaliad comúnmente usan las vocales y las vocales nasales. Sin embargo para obtener una evaluación más general de esta anormalidad es necesario analizar las paradas y las fricativas. Este estudio describe un método con alta capacidad de generalización para detección de hipernasalidad análisis de las consonantes oclusivas sordas españolas. Se muestra la importancia del análisis fonema por fonema, en contraste con la parametrización de la palabra completa que incluye segmentos irrelevantes desde el punto de vista de la clasificación. Los parámetros que correlacionan la incompetencia velofaringea (VPI) sobre las consonantes oclusivas sordas se usa en la fase de estimación de características. La clasificación se llevó a cabo usando una Maquina de Vector de Soporte (SVM), incluyendo el modelo de complejidad Rademacher con el objetivo de aumentar la capacidad de generalización. Rendimientos del 95.2% y del 92.7% fueron obtenidos en las etapas de elaboración y verificación para una repetida evaluación y clasificación de validación cruzada.

Palabras clave

Análisis acústico, análisis del habla, hipernasalidad, consonantes oclusivas sordas y complejidad Rademacher.

1. Introduction

The verbal communication process requires translation of thoughts into spoken language. A person with a physical and/or neurological impairment, may have a compromised vocal tract configuration and/or excitation, resulting in reduced speech quality. A specific example of a vocal tract dysfunction causing reduced speech quality is a defective velopharyngeal mechanism (Cairns et al., 1996). The term cleft palate refers to a malformation which affects the soft and/or hard palate, and it is usually congenital (Vijayalakshmi et al., 2007). Hypernasal analysis based exclusively on hearing is affected by human percepción facts; therefore, the use of measuring tools is important. Digital voice processing (DVP)-based techniques are amongst the most useful of noninvasive techniques for assessing the velopharyngeal function, due to the ease of recording speech signals, which are mainly affected in two ways: 1) nasalized phonemes, and 2) weak consonants and short utterance length (Kummer, 2001). The most common way to detect velopharyngeal disfunction (employing DVP) is by carrying out an analysis of nasalized vowel sounds. In (Vijayalakshmi et al., 2007), a group of delay-based signal processing techniques was described for the analysis and detection of hypernasal speech. Experiments were carried out on the phonemes /*a*/, /*i*/, and /*u*/, where the results showed a high performance on hypernasality detection. The effectiveness of these delay-based acoustic measures were cross-verified on data collected in an entirely different recording environment, however, the generalization capability results of this feature set with regards to the classification accuracy were not convincing. In (Cairns et al., 1996), the sensitivity of the Teager energy operator for multicomponent signals was used for detecting the hypernasality problem. A measurable difference was observed between the low-pass and band-pass profiles for the nasalized vowels, whereas the normal vowel, which is a single component signal, does not show any difference. Parameters such as Harmonics to Noise Ratio (HNR) (Yumoto et al., 1982), Normalized Noise Energy (NNE)

(Kasuya et al., 1986), Glottal to Noise Excitation (GNE) and so on have been proposed for the analysis of pathological voices in different studies. They were mainly designed for sustained vowels, although sometimes they have been used for voiced phonemes, as in (Daza et al., 2008). The real problem in hypernasality detection employing DVP is the high variability within-classes, which means high complexity in the training stage and low generalization capability. In this study, unvoiced consonant analysis is proposed, which impedes the use of features previously developed in the literature for speech pathology assessment. Using parameters that correlate the imprints of Velopharyngeal Incompetence (VPI) over voiceless stop consonants such as power, duration and so on, allow a better representation of the phenomenon currently analyzed. Additionally, finding a reduced representation space of the normal and pathological records is very important, since this procedure reduces computational complexity without loss of classification accuracy and improves the robustness in detection by the Rademacher complexity model, due to the addition of an uncertainty component in the feature subset evaluation stage.

2. Materials and methods

It is necessary to take into account the drawbacks caused by small training samples in the design of automatic classification systems. To reduce these problems, features used must correlate the influence of velopharyngeal incompetence in stop consonants, and classifiers with good generalization properties should be employed (Jain et al., 2000).

2.1. Database

The sample was constituted by 88 children. Classes were balanced (44 patients with normal voice and 44 with hypernasality), and all registers were evaluated by specialists. Each recording was made up by several Spanish words, but in this study only the words coco (*/'koko/*) and *papá* (*/pa'pa/*) were used. Signals were acquired under low noise conditions using a dynamic, unidirectional microphone (cardioide). Signal range was between (-1, 1). A manual segmentation process was carried out to separate the stop parts of the utterances */`'koko/* and */pa'pa/* resulting in various sets (two from */'koko/* and two from */pa'pa/*) each formed by 88 signals.

2.2. Parametrization of plosive signals

A plosive consonant is formed by blocking the oral cavity at some point. During the articulation of most plosives the velum is raised, blocking off the nasal passages. Individuals with cleft palate have never learnt to control the movements of the velum. After reconstructive surgery or the fitting of a prothesis, such individuals need guidance in controlling the velum to produce plosive sounds (MacKay, 1987). The subglottal pressure represents the energy immediately available for creating the acoustic signals of speech (Baken, 1996). The pressure that builds up behind the occlusion is released suddenly as a minor *explosion* or *popping* (Kummer, 2001). The *power* of stops can help to perceive the weakness of plosive consonants in velopharyngeal patients. In this study, it is calculated using the expression:

$$P = 10 \log \frac{P_{stop}}{P_{word}}$$

where

$$P_{\frac{stop}{word}} = \frac{1}{T} \sum_i x_i^2$$

and is the power of the uttered word. Each stop segment is considered separately for the whole database. Air leakage around the blockage significantly slows down the rise in supraglottal pressure, and therefore, delays phonatory shut-down (Baken, 1996). This can provoke a short utterance length of consonant plosives, which in this study, is measured in seconds. Velicaction allows the nasal cavities to be closed or open (or partially open, although air can leak around the velic blockage) with respect to the rest of the vocal tract, which allows sound waves to resonate within the nasal cavities, giving a distinctive nasal quality to the speech sounds produced (MacKay, 1987). In addition, the lower pressure of voiced stops in hypernasal speech results in reduced high-frequency energy for the burst (Baken, 1996). The MFCC (Mel-Frequency Cepstral Coefficients) and DWT (Discrete Wavelet Transform) use filterbanks to obtain measures of different portions of the spectrum, so the energies at every filter could be used to model the behavior at different frequency ranges (Huang et al., 2001). MFCC's are currently one of the most widely used features for Automatic Speech Recognition (Avendano et al., 2005). These features are calculated by taking the discrete cosine transform of the logarithm of energy at the output of a *Mel filter*. In feature estimation processes based on the Fourier transform, the features that are extracted have fixed time frequency resolution because of the inherent limitation of the FFT. More recently, discrete wavelet transform (DWT) and wavelet packets (WP) have been tried for feature extraction, because of their multi-resolution capabilities (Farooq & Datta, 2003).

2.3. Feature selection

In general, given a set of observations represented for a set of features where each observation is associated to one and only one class label from a label set k, the main goal of *feature selection* is to choose the *best* possible subset $\hat{\xi}_i \subseteq \{F_j\}_{j=1}^{p}$ of size q from a set of p features, where optimal and suboptimal strategies are ussually considered. For the optimal case, if the cardinal of $\hat{\xi}_i$ is q, and all the q-cardinal subsets are in $\breve{\xi}$, the subset $\hat{\xi}_i$ is that which optimizes a evaluation function f, such as (Jain et al., 2000):

$$f\left(k,\hat{\xi}_i\right)=\max_{\breve{\xi}\subseteq\{F_j\}} f\left(k,\breve{\xi}\right)$$

In pattern recognition tasks, feature selection according to the evaluation function f can be carried out by *wrapper type selection*, when f uses information of the classification function oriented to minimizing the classifier error, and *filter type selection*, which consists in data preprocessing by optimizing f with respect to a metric (independent of the classification results), where the irrelevant, redundant and correlated variables are discarded (Webb, 2002). Wrapper type selection procedures give better performance in cases when the number of features is lower than 50 (Kudo, & Sklansky, 2000), while the filter type can operate in larger spaces because its computational demand usually is lower (Jain et al., 2000). Suboptimal algorithms, although incapable of examining every feature combination, will assess a set of potentially useful feature combinations. Popular methods such as sequential forward selection (SFS) and sequential backward selection (SBS) are found. In floating search methods such as sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS), the number of added and removed features can change at each step and these wrapper routines carry out the search in a considerably smaller number of subsets (Alpaydin, 2004).

2.4. Support vector classifiers

Support Vector Machines (SVMs) were used in this study mainly for two reasons: SVMs have a relatively good generalization capability with less amount of training data, and they have been particularly well developed for binary classification tasks. Traditional neural network approaches are more likely to suffer of poor generalization, producing models that can overfit the data. It is a consequence of the optimization algorithms used for parameter selection and the statistical measures used to select the *best* model (Solera et al., 2007). For the binary classification problem, a discrimination function can be taken as

$$g\,(x) = w^T\,\varnothing\,(x) + w_0$$

with decision rules $w^T \varnothing\,(x) + w_0 \geq 0 \rightarrow x \in w_1$ and $w^T \varnothing\,(x) + w_0 \leq 0 \rightarrow x \in w_2$, where

$\phi(x):\Re^{n1}\rightarrow\Re^{n2}$ is generally a nonlinear function which maps vector x into what is called a feature space of higher dimensionality (possibly infinite) where classes are linearly separable. The vector w defines the separating hyper-plane in such a space and w0 represents a possible bias (Webb, 2002).

2.5. Rademacher complexity model

Rademacher complexity is a measure proposed in (Koltchinskii, 2001) which attempts to balance the complexity of the model with its fit to the data by minimizing the sum of the training error and a penalty term. Let $\{X_i, Y_i\}^n_{i=1}$ be a set of training instances, where X_i is the pattern or example associated with features $\{F_j\}^q_{j=1}$, and Y_i is the label of the example X_i. Let $h(x_i)$ be the class obtained by the classifier h, trained using $\{X_i, Y_i\}^n_{i=1}$. Then, the training error is defined as $\hat{e}(h)=\frac{1}{n}\sum_{i=1}^{n} I_{\{h(x_1)\neq y_1\}}$ where,

$$I_{\{h(x_1)\neq y_1\}}=\begin{cases} 1, \text{ when } h(X_i)\neq Y_i \\ 0, \text{ when } h(X_i)=Y_i \end{cases}$$

Let $\{\sigma_i\}^n_{i=1}$ be a sequence of Rademacher random variables (i.i.d.) independent of the data $\{X_i\}^n_{i=1}$ and each variable takes values +1 and -1 with probability 1/2. According to this, computation of the Rademacher complexity involves the following steps (Delgado et al., 2007):

– Generate $\{\sigma_i\}^n_{i=1}$

– Get a new set of labels, doing $z_i = \sigma_i y_i$.

– Train the classifier h_R using $\{X_i, Z_i\}^n_{i=1}$.

– Compute the Rademacher penalty, given by

$$\mathrm{R_n}=\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i I_{\{h_R\neq Y_i\}}\right|$$

– Train the classifier h, using $\{X_i, Y_i\}^n{}_{i=1}$

– Compute the training error $\hat{e}(h)$.

– The Rademacher complexity: $RC = \hat{e}\ (h) + R_n$

2.6. Proposed procedure

The representation space was composed of the features related to the plosive consonant power and its duration with respect to the word. On the other hand, feature estimation in the frequency domain was achieved by using two techniques: DWT and MFCC. Each feature was estimated for each plosive consonant at the beginning of */kóko/* and */papá/*. By using 3rd order spline mother wavelet, the Nyquist spectral range was divided in 4 bands (i.e., 3 for the detail levels and 1 for the approximation level). The other features consist of estimating parameters related to 13-MFCC. With the aim of comparing these two representation forms with regards to the discriminant capability, the classification results were obtained using a SVM classifier. Thus, the total number of extracted features for each observation was 15. Feature space reduction is carried out by using a typewrapper algorithm for heuristic search (i.e., SFFS algorithm) with a SVM classifier and RBF-kernel (a kernel successfully used in several speech-related applications) using a hypothesis test based on a distance measurement for establishing the initial conditions. Moreover, the Rademacher complexity model has been included in the evaluation function f. With the aim of comparing the proposed model's performance, the conventional training was developed and proved under the same conditions.

3. Results and discussion

The utterance */'koko/* has two plosive segments, Figures 1 and 2 show 2-dimensional scatter plots using the duration and power for each segment.

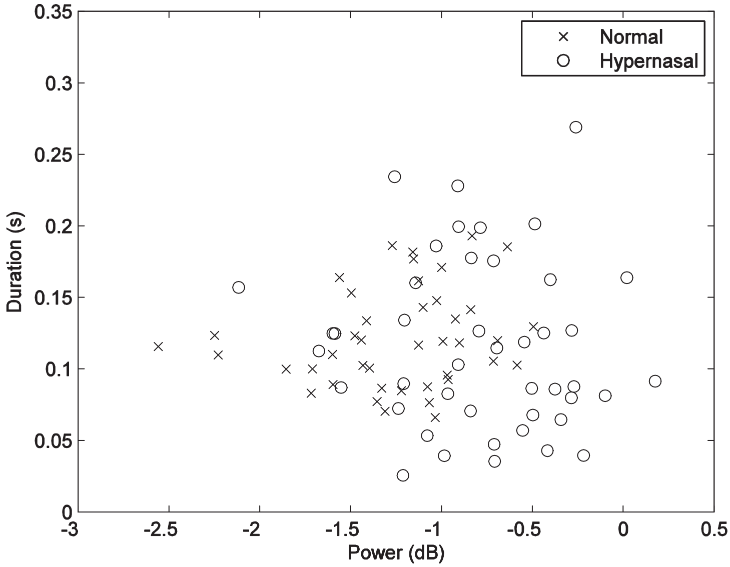

Figure 1: Duration vs power for the first plosive segment in the Spanish word /'koko/

Discrimination between the two classes can be observed using the first plosive segment. By contrast, in the second figure this configuration can not be seen. The closure of the velopharyngeal gap is necessary to produce vowels as well as stops; but in the first segment the velopharyngeral gap begins to open, provoking a delay in the closing phase, in hypernasal speech. At the beginning of the second stop, the velum is closed as the previous phoneme is a vowel, thus the duration is more similar to the normal category, as depicted in Figure 2. Other parameters used in this study are related to the fact that spectral components of the plosives are modified by the velopharyngeal incompetence. This set is formed by the energy for every band in a 3rd level of the DWT decomposition. In Figures 3 and 4 two of these values are shown. Once again there is greater similarity between the two classes in the second stop segment.

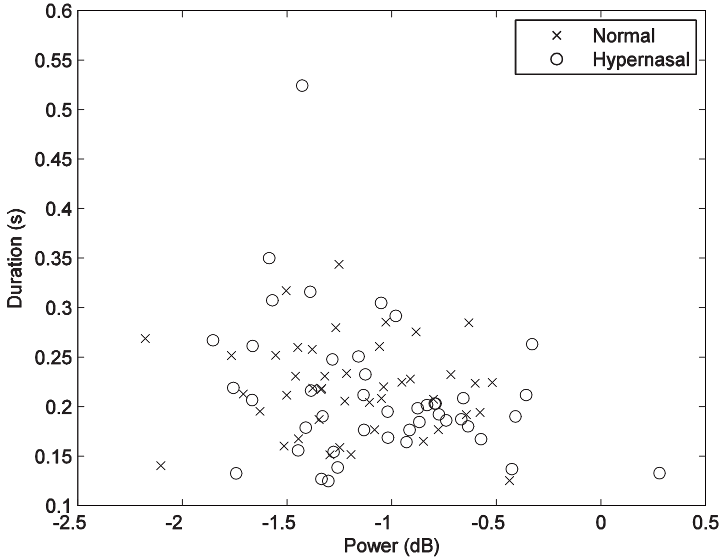

FIGURE 2: DURATION VS POWER FOR THE SECOND PLOSIVE SEGMENT IN THE SPANISH WORD /'KOKO/

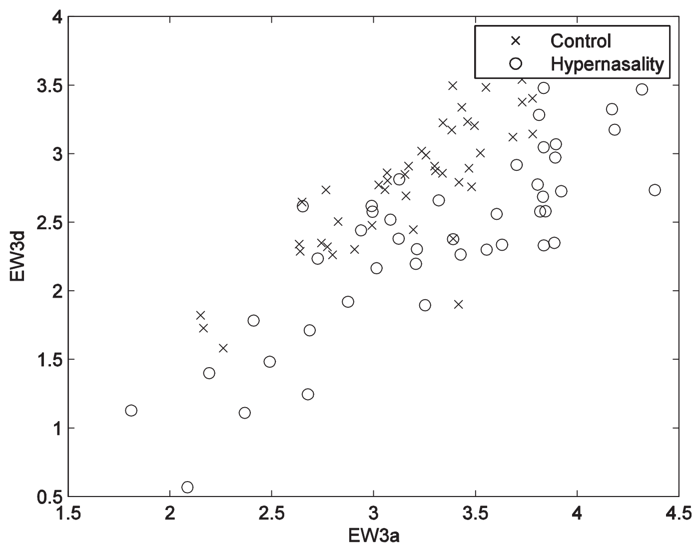

FIGURE 3: ENERGY IN THE THIRD APPROXIMATION AND DETAIL BANDS FOR THE FIRST PLOSIVE SEGMENT IN THE SPANISH WORD /'KOKO/

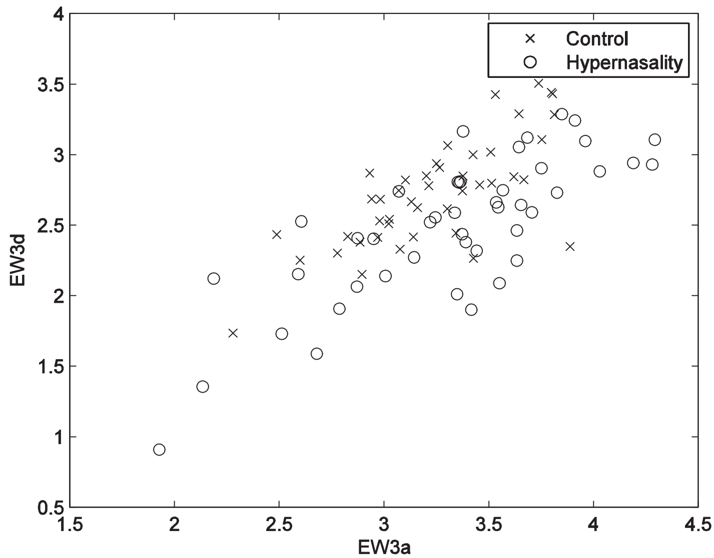

Figure 4: Energy in the third approximation and detail bands for the second plosive segment in the Spanish word /'koko/

The two sets are slightly distinguishable as can be seen in the energy-band parameter distribution, nevertheless when these parameters were evaluated (joined with *Power* and *Duration*) from the point of view of the classification rate, the performance does not reach 61%. When 13th order MFCC coefficients were calculated, instead of DWT, the performance went up to levels between 83% and 88% with an average rate of 85.7%, although this result fell down to 62.4% in the verification stage (i.e., poor generalization capability). The classifier evaluation was made by applying cross-validation for 30 folds. Similar experiments were carried out for each word as it is shown in Table 1, where F1p,v is the feature set related to power and duration + DWT for the processing and verification data, similarly, F2p,v is the feature set related to power and duration + MFCC. The notation // + // means that in this case the whole feature set for */'koko/* and */pa'pa/* has been considered. Feature selection results obtained without/with the rademacher complexity model included in the evaluation function are shown in Table 2, over the processing and verification data sets, where $\#F_{\Re,R}$ is the selected feature number from the $\{F1+F2\}$ feature set without/with the rademacher complexity model, similarly, $A^{p,v}_{\Re,R}$ is the average classification accuracy (%) for the processing and verification data. It is remarkable that the classification results for the processing with the Rademacher complexity model are lower than for the other cases, although finally in the verification stage, the proposed training retains the classification accuracy even when the input samples are completely unknown.

TABLE1: CLASSIFICATION RESULTS (%) FOR /KÓKO/ AND /PAPÁ/

	$F1^p$	$F2^p$	$F1^v$	$F2^v$
/'koko/ /pa'pa/	59.8 60.9	92.8 97.3	52.1 55.3	63.7 77.9
//+//	85.7		62.4	

TABLE 2: RESULTS WITHOUT/WITH THE RADEMACHER COMPLEXITY

	$\#F_{\Re}$	$A^p_{\Re}$	$A^v_{\Re}$	$\#F_R$	A^p_R	A^v_R
/'koko/ /pa'pa/	5 6	89.6 95.1	62.8 78.3	9 7	87.6 93.2	85.2 89.5
//+//	7 96.6		76.8 12		**95.2**	**92.7**

4. CONCLUSIONS

From these experiments it can be concluded that hypernasal assessment should be determined analyzing phoneme by phoneme, instead of complete words. The acoustic properties of the same phoneme can be completely different in different parts of the uttered word due to variability of the behavior of articulators which depend so much on the context. The obtained results show that the Rademacher penalty adds generalization capability to the classifier, which is a necessary constraint due to the high within-class variability of speech signals. This uncertainty included in the feature selection allows effective dimensionality reduction. Using few features, a performance of 85.2% in the verification stage was obtained for the voiceless plosive */k/*, 89.5% for the phoneme */p/* and 92.7% considering both phonemes. Thus, feature selection revealed what features contributed to the generalization capability. For example, the *power* has discriminant information for */p/*, while the phoneme */k/* is well-represented by the *duration*. The other selected features were related to the high-frequency bands except one feature of low-frequency, which could probably be used as reference for the classifier. This is in agreement with the information supplied by the clinic specialists (Baken, 1996).

5. Acknowledgements

This study is framed within the project titled “Identificación de posturas labiales en pacientes con labio y/o paladar hendido corregido”, financed by COLCIENCIAS.

6. References

Alpaydin, E. (2004). Introduction to Machine Learning. MIT Press.

Avendano, A., Deng, L., Hermansky, H., & Gold, B. (2005). The Analysis and Representation of Speech, in Speech Processing in the Auditory System, Springer Verlag.

Baken, R. J. (1996). Clinical Measurement of Speech and Voice. Singular Publishing Group, Inc.

Cairns, D., Hansen, J., & Kaiser, J. (1996). Recent advances in hypernasal speech detection using the nonlinear teager energy operator. International Conference on Spoken Language Processing, vol. 2, pp. 780–783.

Daza, G., Sánchez, L., Sepúlveda, A., & Castellanos, G. (2008). Acoustic feature analysis for hypernasality detection in children. Encyclopaedia of Healthcare Information Systems. IDEA Group, Inc.

Delgado, E., Giraldo, L. F., & Castellanos, G. (2007). Feature selection using a hybrid approach based on the rademacher complexity model selection. Computers in Cardiology, vol. 34, pp. 257–260.

Farooq, O., & Datta, S. (2003) Phoneme recognition using wavelet based features. Information Sciences, vol. 150, pp. 5–15.

Huang, X., Acero, A., & Hon, H. W. (2001). Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall.

Jain, A., Duin, R., & Mao, J. (2000). Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4–36.

Kasuya, H., Ogawa, S., Mashima, K., & Ebihara, S. (1986). Normalized noise energy as an acoustic measure to evaluate pathologic voice. Journal of Acoustical Society of America, vol. 80, no. 5, pp. 1329– 34.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. IEEE Transactions on Information Theory, vol. 47, no. 2, pp. 1902–1914.

Kudo, M., & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. Pattern Recognition, vol. 33, pp. 25–41.

Kummer, A.(2001). Cleft Palate and Craniofacial Anomalies: Effects on-Speech and Resonance. Thomson Delmar Learning.

MacKay, I. R. (1987). Phonetics: the science of speech production. Allyn and Bacon.

Solera, R., Martín, D., Gallardo, A., Pelaéz, C., & Díaz, F. (2007). Robust asr using support vector machines. Speech Comunication, vol. 49, pp. 253–267.

Vijayalakshmi, P., Ramasubba, M., & O'Shaughnessy, D. (2007). Acoustic analysis and detection of hypernasality using a group delay function. IEEE Transactions on Biomedical Engineering, vol. 54, no. 4, pp. 621–629.

Webb, A. R. (2002). Statistical Pattern Recognition. John Wiley and Sons, Ltda.

Yumoto, E., Gould, W., & Baer, T. (1982). Harmonics-to-noise as an index of the degree of hoarseness. Journal of Acoustical Society of America, vol. 71, no. 6, pp. 1544–1550.