DOI: http:/dx.doi.org/10.18180/tecciencia.2015.19.3

TECCIENCIA

Phoneme Recognition System Using Articulatory-Type Information

Sistema de reconocimiento de fonemas usando información de tipo articulatorio

Alberto Patiño Saucedo[1*], Alexander Sepúlveda Sepúlveda[2], Diego Ferney Gómez Cajas[3]

[1] *Universidad Industrial de Santander, Bucaramanga, Colombia, alberto.patino@correo.uis.edu.co*
[2]*Universidad Industrial de Santander.Bucaramanga, Colombia, fasepul@uis.edu.co*
[3]*Universidad Antonio Nariño, Bogotá, Colombia, dfgomezc@uan.edu.co*

Received: 05 April 2015 *Accepted: 18 Jun 2015* *Available Online: 23 June 2015*

Abstract

This work is frameworked within the development of phoneme recognition systems and seeks to establish whether the incorporation of information related to the movement of the articulators helps to improve the performance thereof. For this purpose, a pair of systems is compared and developed, where the acoustic model is obtained from training hidden Markov chains. The first system represents the voice signal by Mel Frequency Cepstral Coefficients; the second uses the same Cepstral coefficients but together with articulatory parameters. The experiments were conducted on the MOCHA-TIMIT database. The results show a significant increase in the system´s performance by adding articulatory parameters compared to that based only on Mel Frequency Cepstral Coefficients.

Keywords: Mel-Cepstrum coefficients, hidden Markov models, articulatory parameters, phoneme recognition

Resumen

El presente trabajo se enmarca dentro del desarrollo de sistemas de reconocimiento de fonemas y busca establecer si la incorporación de información relacionada con el movimiento de los articuladores ayuda a mejorar el desempeño de los mismos. Para ello, se desarrollan y comparan un par de sistemas, donde el modelo acústico se obtiene a partir del entrenamiento de cadenas ocultas de Markov. El primer sistema representa la señal de voz mediante coeficientes cepstrales en la escala Mel; y el segundo, utiliza los mismos coeficientes cepstrales pero en conjunto con parámetros articulatorios. Los experimentos fueron realizados sobre la base de datos MOCHA-TIMIT. Los resultados muestran un incremento significativo en el desempeño del sistema al agregar parámetros articulatorios con respecto a sistemas basados en coeficientes cepstrales en la escala Mel.

***Palabras clave*:** coeficientes cepstrales en la escala de Mel, modelos ocultos de Markov, parámetros articulatorios, reconocimiento de fonemas

1. Introduction

Automatic speech recognition has been the object of intense research for over four decades, reaching notable results. However, while tasks like digit recognition have reached rates of 99.6% [1]e, the same cannot be said for phoneme recognition for which the best rates are around 80% [2]. Phoneme recognition is a relevant task that could improve the performance of other voice signal processing systems like the very automatic speech recognition (ASR) [3], speaker verification [4], and learning of a second language [5], among others. The capacity to recognize phonemes with high precision becomes, then, a fundamental problem in the field of language processing. One of the advantages of phoneme recognition is its versatility, given that it permits knowing the phonetic characteristics of speech, independent of the vocabulary, adapting to different languages, in contrast to speech recognition focused on words or phrases.

To train phoneme recognition systems we must have speech databases with their respective phonemes adequately segmented and validated, which requires costly and intensive processes in skilled labor. Within those databases, two are highlighted: TIMIT and MOCHA [6].

For example, in [7] the TIMIT database is used with a strategy that consists in using cascaded neural networks for a phoneme recognition task independent of the speaker. During the first stage, the probability values of the phonemes are estimated; and during the second stage, these estimated values are used to make up the vector of

***Corresponding Author.**
E-mail: *alberto.patino@correo.uis.edu.co*

How to cite: Patiño, A.; Sepulveda, A.; Gomez, D., ***Phoneme Recognition System Using Articulatory-Type Information*,** TECCIENCIA, Vol. 7 No. 19., 11-14**,** 2015, **DOI:** http:/dx.doi.org/10.18180/tecciencia.2015.19.3

characteristics of the recognition system. With the aforementioned, a 72% rate of success is accomplished using context windows of 230 ms for the second stage of neural networks. Similarly, [8] uses a layer of neural networks to estimate probabilities *a posteriori* of the phonemes and a second layer to classify the phonemes from the probabilities estimated, achieving a phoneme error rate (PER) of 19.6%. An additional strategy, denominated deep belief networks (DBN), is exposed in [2]. Also, for phoneme recognition hidden Markov models (HMM) have also been used [9] along with classifiers based on Support Vector Machines (SVM) [10]. Nevertheless, to represent the speech signal, conventional characteristics are used, like Mel-Frequency Cepstral Coefficients (MFCC) and perceptual linear predictive (PLP) analysis for which performance drops as noise starts to affect the speech signal [11].

One of the ways to improve the performance of phoneme recognition systems consists in using alternative representations to the classical representations exposed. For example, another work [12] uses myoelectric-type signals, as complement of the speech signal, in a phoneme recognition system based on hidden Markov models. Said work shows that the system's performance improves, particularly under noise conditions. In addition, other authors [13] use information extracted from acoustic waves travelling through the body tissue of people when speaking, whose signals are picked up by special microphones placed behind the ear. Likewise, several methods have been developed seeking to include visual information about lip movement to improve recognition systems. An alternative representation to the ways previously described corresponds to the use of information on the movement of articulators instead of using only information of the acoustic signal. This type of information can be obtained through devices capable of measuring the movement of the articulators in the vocal tract. Among these devices, we can highlight the recent development of the Electro-Magnetic Articulograph (EMA) [14].

Given the aforementioned, this work sought to establish if incorporating articulatory data can improve the rate of phoneme recognition, comparing the performance of a classical recognition system (based on hidden Markov models and Cepstral coefficients) to another in which besides using MFCC coefficients, information of articulatory nature is used.

2. Materials and methods

2.1. Database

The articulatory data in this work were obtained from the MOCHA database, given that it provides phonetically diverse voice signals (desirable for the training task). This database also includes four data sequences recorded simultaneously: the acoustic signal with a sampling frequency of 16 KHz, laryngography, electropalatography, and EMA data. In relation to EMA data, sensors are installed in the lower incisors (li), the upper lip (up), the lower lip (ll), the tip of the tongue (tt), tongue body (tb), tongue dorsum (td), and soft palate (v). The two sensors on the bridge of the nose and the upper incisors provide points of reference that permit correcting the errors produced by the head movements. The EMA trajectories are sub-sampled at 100 Hz, after an anti-aliasing filtering process. Thereafter, given that the movements of the articulators generally have bandwidths below 15 Hz, the EMA trajectories are softened with a low-pass filter whose cutoff frequency is 20 Hz. The filtering process of EMA signals is carried out directly and inversely to avoid possible phase distortions over the signal [15].

As per data standardization, a process suggested in [16] was carried out. The conventional process calculates the average and global standard deviation values to then apply them to the EMA trajectories; but this could generate difficulties because the mean values vary from one phrase to another during the recording process. Also, it is worth highlighting that while the rapid changes of the mean values are attributed to the phonetic content of each phrase, the slow changes are mainly caused by the subject's articulatory adaptation during the recording session. Hence, it turns out useful to eliminate the first class of variation, maintaining the second. This is accomplished by subtracting a version of mean values obtained by passing the mean vector through a low-pass filter [15].

2.2. Representation

Regarding signal representation, this work uses two types: the first corresponds to parameters that describe the behavior of the acoustic signal; the second corresponds to the position of the articulating organs with respect to a particular point of reference. In relation to the representation of the acoustic signal, MFCC coefficients are used [17]. This are widely used to obtain the characteristic vectors of the speech signal. Said technique is inspired on the functioning of the most important organ intervening in human hearing: the cochlea.

To obtain the MFCC, the speech signal is first filtered through a pre-emphasis single-zero high-pass filter, located at 0.97. Then, a window processes is conducted by selecting 25-ms lengths of the signal, at a rate of 100 Hz (every 10 ms). Each 25-ms block was applied the following procedure:

1. The discrete-time Fourier transform was calculated and the magnitude response in frequency was obtained. The phase was not used, given that, generally, the human ear does not distinguish small phase variations.

2. A bank of 12 filters was applied to the spectrum's magnitude response and the denominated Energy Bands from each filter were obtained [11].

3. The logarithm was applied to the energies of the filter bank to construct a vector of R length.
4. The discrete cosine transform (DCT) was calculated to the previous vector.

The bank of M triangular filters is defined thus [18]:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{2(k - f(m-1))}{(f(m+1) - f(m-1))(f(m) - f(m-1))}, & f(m-1) < k < f(m) \\ \dfrac{2(f(m+1) - k)}{(f(m+1) - f(m-1))(f(m+1) - f(m))}, & f(m) < k < f(m+1) \\ 0, & k > f(m+1) \end{cases} \quad (1)$$

Where the $f(m)$ values are given by the expression:

$$f(m) = \frac{N}{F_s} B^{-1}(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}) \quad (2)$$

With f_l and f_h being the minimum and maximum frequencies of the bank of filters in Hz, F_s the sampling frequency, and N the size of the Fourier discrete transform of the speech signal portion. B is the function that approximates the frequency values in the Mel scale.

$$B(f) = 1125 \ln(1 + \frac{f}{700}) \quad (3)$$

B^{-1} corresponds to the inverse of the approximation function. **Figure 1** shows a bank of 12 filters. Note that the bandwidth ratio is proportional to the central frequency. For each filter, the power sum is calculated.

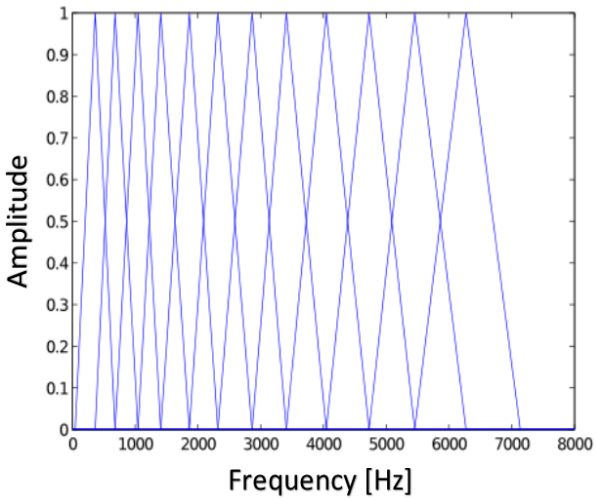

Figure 1 Bank of 12 MFCC filters

2.3. Modeling

Speech recognition systems consist of a series of statistical models that represent the different sounds to be recognized, in this case the phonemes.

By the speech signal having a temporary structure it may be encoded as a sequence of spectral vectors: the MFCC coefficients. Due to this circumstance, the hidden Markov models (HMMs) can be used to construct said models from the characteristic vectors of the speech signal [19].

A Markov model is a finite state machine that changes its state every given unit of time, where it is assumed that the observations sequence: $\boldsymbol{O} = \boldsymbol{o_1}, \boldsymbol{o_2}, \ldots, \boldsymbol{o_T}$, is generated by this state machine (**Figure 2**).

Every time, t, in which a j state is input, a characteristic vector $\boldsymbol{o_t}$ is generated, according to the probability density $b_j(\boldsymbol{o_t})$. Additionally, the transition from state i to j is also probabilistic and is given by the discrete probability a_{ij}. In practice, only the observation sequence $\boldsymbol{O}$ is known, while the underlying sequence of states is unknown, although it may be calculated by using equation:

$$P(\boldsymbol{O}|M) = \sum_{X} \alpha_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(\boldsymbol{o_t}) a_{x(t)x(t-1)} \quad (4)$$

Where $\boldsymbol{X} = x(1), x(2), x(3), \ldots, x(t)$, is the vector of the model's possible states, with $x(0)$ being the initial state and $x(T+1)$ the output state. Although, the direct calculation of this equation is not performed, recursive methods exist that permit for both amounts to be calculated efficiently, like the Baum-Welch and the Viterbi algorithms [19].

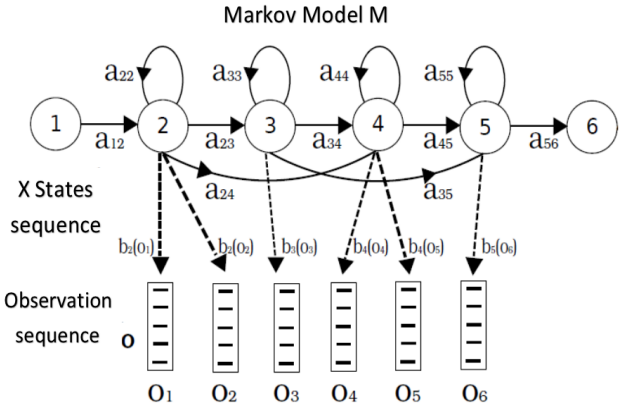

Figure 2 Hidden Markov models represented as a state machine. Adapted from the HTK-Book **[20]**.

2.4. *Experiment configuration*

For this stage of the study, HTK software was used for the task of extraction of characteristics and for modeling with HMMs and its corresponding training and recognition stages. As mentioned in the previous section, in the first instance a pre-emphasis FIR filter was applied (with coefficients 1 and -0.97) and a Hamming window with a sampling frequency of 100 Hz (equal to the EMA data) whose size is 25 ms per window. Thereafter, the 22 filters were obtained in the Mel scale from which 13 MFCC coefficients were generated, with their respective delta and delta-delta coefficients with which a 39-element vector was obtained for each speech block. For the case when adding articulatory information, the 14 components from the articulatory vector were added to this vector, resulting in a 53-element vector. The HMMs were initialized by using the Viterbi algorithm and then the model parameters were estimated by using the Baum-Welch algorithm.

The experimentation proposed analyzes four systems developed: two with articulatory data (for both speakers: male and female), and two similar without articulatory data. To evaluate the system performance, the data available are separated into the training set and the test data set. The MOCHA-TIMIT database has 460 phrases labeled in approximately 16,000 phonemes for two speakers from both sexes. A total of 80% of the phrases is used for training and the remaining 20% is used for validation, as done in the system implemented in [21]; where the phrases are selected randomly without replacement. In total, 15 experiments are carried out for each system.

3. Results and discussion

One of the most broadly used ways to evaluate phoneme recognition systems is the phonetic error rate (PER) [22] [18]. It measures the difference between the sequences of recognized phonemes with the correct sequence and is calculated by adding the total of errors over the number of phonemes of the correct sequence (N). The aforementioned is expressed with formula:

$$PER = 100 * \frac{I + S + D}{N} \quad (5)$$

Where, *I, S, D* are defined by other works [18]: *S* (errors by substitution), when an incorrect phoneme substitutes a correct one; *D* (errors by omission), when a correct phoneme is omitted; and *I* (errors by insertion), when an extra phoneme is added. The precision value (*A*) is calculated from the PER in the following manner:

$$A = 100 - PER \quad (6)$$

Another way of measuring performance consists in using the amount of phonemes correctly identified by the system, which we will call rate of success C. This index (C) is calculated without bearing in mind errors by insertion in the following way:

$$C = 100 * \frac{N - S - D}{N} \quad (7)$$

The phoneme recognition process is carried out 15 times to generate a vector with the performance values; where, for each of these times the voice registries of the test and training sets were randomly chosen without replacement. This is for the system based on MFCC and in MFCC+AP (articulatory parameters). **Figure 3** and **Figure 4** show the precision and success rates, respectively, for speakers fsew0 and msak0.

This work tests the hypothesis that proposes that using articulatory parameters helps to improve the performance of phoneme recognition systems. For this, we use the Student t statistical test for unknown means and variances. With the Student t it can be verified if the mean values of the populations are significantly removed; that is, if the performance improvement is significant. As a result, it is found that the hypothesis test is fulfilled; thereby, it is said that it cannot be dismissed that the difference is significant. In other words, it is completely reasonable to state that performance improves significantly. Improvements in recognition rates calculated are summarized in **Table 1.**

Table 1 Improvements in precision rates and correct phoneme percentage for both speakers

	fsew0	**msak0**
A	11.17%	10.53%
C	5.71%	7.03%

The highest precision rate obtained was 69.26% for the second test of the combined system (speaker *msak0*). It may be seen from graphics 3 and 4 that the variance of the results is small, both for the male and female speakers; besides, the results between the MFCC and MFCC+EMA sets never overlap each other. This indicates that the system is reliable and that its performance improvement is not due to chance.

Regarding the precision rate (A), it is observed that the difference between using MFCC and MFCC+EMA sets is notable; which supports the importance of using articulatory information in phoneme recognition tasks. Furthermore, referring to percentage of correct phonemes (C), the difference is also noted at plain site. The Student t tests corroborate the prior statements; from which it is inferred that the population means corresponding to the performance of the MFCC and MFCC+EMA sets are statistically significant.

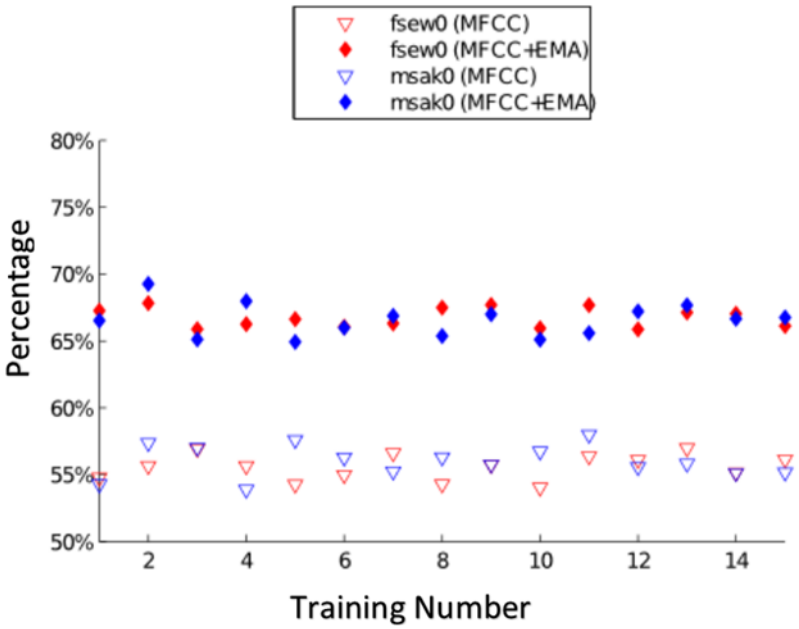

Figure 3 Results of the precision rate (A) for both speakers

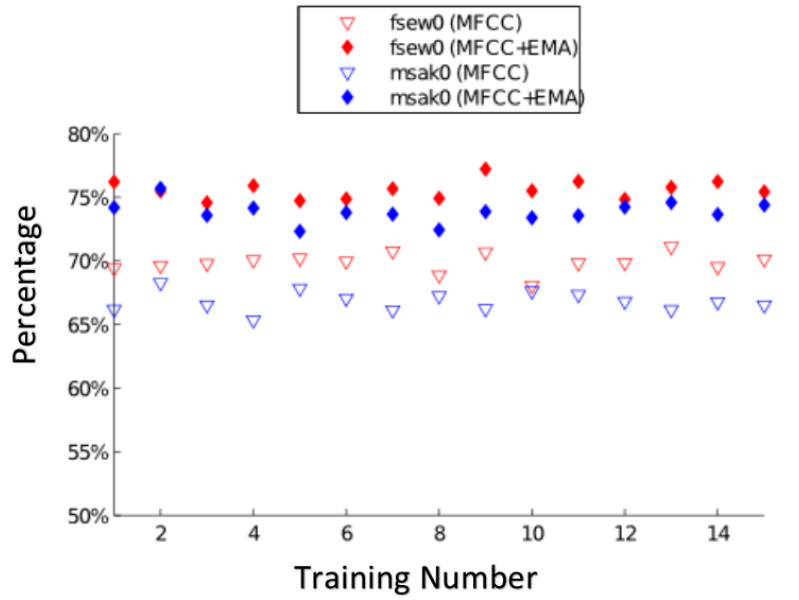

Figure 4 . Results of correct phonemes over the total (C) for both speakers

4. Conclusions

This work showed that incorporating articulatory parameters, as voice representation, can improve the rate of phoneme recognition based on hidden Markov chains. The results exhibited a significant increase in the system's performance upon adding articulatory parameters with respect to a system based on MFCC parameters, which leads to inferring that the articulatory parameters provide relevant information for phoneme recognition purposes. In addition, it is worth highlighting that the possibility remains open to link articulatory information to automatic continuous speaking recognition systems to analyze the positive effect it may insert on the system's performance

References

[1] J. Li, «Soft Margin estimation for automatic speech recognition,» Georgia Institute of technology- Dissertation presented to the academic faculty, Georgia, 2008.

[2] A.-. r. Mohamed, G. E. Dahl y G. Hinton, «Acoustic modeling using deep belief networks,» *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, nº 1, pp. 14-22, 2012.

[3] K. F. Lee y H. W. Hon, «Speaker-independent phone recognition using hidden Markov models,» *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 37, nº 11, pp. 1641-1648, 1989.

[4] S. Furui, «40 Years of Progress in Automatic Speaker Recognition,» *Lecture Notes in Computer Science,* vol. 5558, pp. 1050-1059, 2009.

[5] S. M. Witt y S. J. Young, «Phone-level pronunciation scoring and assessment for interactive language learning,» *Speech Communication,* vol. 30, nº 2-3, pp. 95-108, 2000.

[6] A. A. Wrench, «A multi-channel/multi-speaker articulatory database for continuous speech recognition research,» *Phonus 5,* pp. 1-13, 2000.

[7] J. Pinto, S. Garimella, M.-. Doss, H. Hermansky y H. Bourlard, «Analysis of MLP-Based Hierarchical Phoneme Posterior Probability Estimator,» *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, nº 2, pp. 225-241, 2011.

[8] G. S. Sivaram y m. H. Hermansky, «Sparse Multilayer Perceptron for Phoneme Recognition,» *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, nº 1, pp. 21-29, 2012.

[9] J. Park y H. Ko, «Real-Time Continuous Phoneme Recognition System Using Class-Dependent Tied-Mixture HMM With HBT Structure for Speech-Driven Lip-Sync,» *IEEE Transactions on Multimedia,* vol. 10, nº 7, pp. 1299-1306, 2008.

[10] J. Yousafzai, P. Sollich, Z. Cveltovic y B. Yu, «Combined features and kernel design for noise robust phoneme classification using support vector machines,» *IEEE Transactions on Audio, Speech, and Language Processing.,* vol. 19, nº 5, pp. 1396-1407, 2011.

[11] D. F. Gomez, A. Sepulveda y M. A. Pinto, «Parametrizaciones robustas de Reconocimiento Automático de Habla (RAH) en redes de comunicaciones,» *Ingeuan,* vol. 3, nº 6, 2013.

[12] E. J. Scheme, B. Hudgins y P. A. Parker, «Myoelectric Signal Classification for Phoneme-Based Speech Recognition,» *IEEE Transactions on Biomedical Engineering,* vol. 54, nº 4, pp. 694-699, 2007.

[13] P. Heracleous, V. A. Tran, T. Nagai y K. Shikano, «Analysis and Recognition of NAM Speech Using HMM Distances and Visual Information,» *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, nº 6, pp. 1528-1538, 2010.

[14] Z. Zhou, G. Zhao, X. Hong y M. Pietikainen, «A review of recent advances in visual speech decoding,» *Image and vision computing,* vol. 32, nº 9, pp. 590-605, 2014.

[15] A. Sepulveda, R. Capobianco Guido y G. Castellanos Dominguez, «Estimation of relevant time–frequency features using Kendall coefficient for articulator position inference,» *Speech communication,* vol. 55, nº 1, pp. 99-110, 2013.

[16] K. Richmond, «Estimating Articulatory parameters from the acoustic speech sign,» PhD thesis, the centre for speech technology research, Edinburgh University, Edinburg, 2002.

[17] S. Davis y P. Mermelstein, «Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,» *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 28, nº 4, pp. 357-366, 1980.

[18] X. Huang, A. Acero y H.-W. Hon, Spoken Language Processing: A guide to theory, algorithm, and system development, Saddle River: Prentice Hall PTR, 2001.

[19] M. Gales y S. Young, «The Application of Hidden Markov Models in Speech Recognition,» *Foundations and trends in signal processing,* vol. 1, nº 3, pp. 195-304, 2008.

[20] S. Young, M. Gales, T. Hain, X. Liu, D. Kershaw y G. Evermann, The HTK Book. Revised for HTK Version 3.4, Cambridge University Engineering Department, 2006.

[21] A. Wrench y K. Richmond, Continuous Speech Recognition Using Articulatory Data, Edinburg: Department of speech and language sciences, Queen margarate university College, 2000.

[22] M. Antal, «Toward a simple phoneme based speech recognition system,» *Informatica,* vol. 52, nº 2, pp. 33-48, 2007.