Representing Phonological features through a two-level finite state model

Javier M. Olaso, María Inés Torres, Raquel Justo

Departamento de Electricidad y Electrónica, Universidad del País Vasco, Spain

javiermikel.olaso@ehu.es, manes.torres@ehu.es, raquel.justo@ehu.es

Abstract

Articulatory information has demonstrated to be useful to improve phone recognition performance in ASR systems, being the use of Neural Networks the most successful method to detect articulatory gestures from the speech signal. On the other hand, Stochastic Finite State Automata (SFSA) have been effectively used in many speech-input natural language tasks. In this work SFSA are used to represent phonological features. A hierarchical model able to consider sequences of acoustic observations along with sequences of phonological features is defined. From this formulation a classifier of articulatory features has been derived and then evaluated over a Spanish phonetic corpus. Experimental results show that this is a promising framework to detect and include phonological knowledge into ASR systems.
Index Terms: phonological features, finite-state models, ASR.

1. Introduction

Phonological feature space has been proposed to represent acoustic models for automatic speech recognition (ASR) tasks. This proposal is based on the speech production mechanism which can be described as a composition of articulatory gestures [1]. The movements of the nearly independent articulators suggests alternatives to the classical segmental models, i.e. phone-like models. Articulatory information has demonstrated to be useful to improve ASR, as reported by many authors [2] [3]. Moreover, the articulatory-based space seems to be more robust for noisy or spontaneous speech and to be less variable in general [4]. In this context, two main goals have been addressed in the last years: the obtention of distinctive phonological features and the use of them to improve ASR system performance.

The most successful method to detect articulatory gestures from the speech signal is based on Time Delay Neural Networks (TDNN) [4]. Alternatively better recognition of articulatory features has been reported using dynamic Bayesian networks [5]. However, acoustic models based on phonetic features does not directly improve phonetic decoding performance [6]. This is probably due to the fact that the segmental models are not fully avoided since phonetic features are derived and then used in a phone-synchronous manner [5] [3]. But significant improvements have been reported when combined with Hidden Markov models (HMM) defined over mel frequency cepstrum coefficients (MFCC) [3] and [7]. On the other hand, the phonological feature space has demonstrated to be useful for lattice rescoring with knowledge scores in ASR tasks, which has lead to significant reductions in phone error rates [8].

Stochastic Finite State Automata (SFSA) have been successfully used when dealing with speech processing tasks. This success is probably due to their ability to include more structural information than pure statistical models. Well established algorithms can also be found to estimate the model parameters. Moreover, they allow an easy combination with other models, which is especially convenient for ASR or speech translation tasks. In particular, stochastic k-Testable in the Strict Sense (k-TSS) models, which can be considered as a syntactic approach of the n-grams models [9] have been proposed for ASR [10], language identification [11], language modeling [12] or machine translation [13].

We propose in this paper the use of k-TSS models to represent phonological features. A hierarchical model able to consider sequences of acoustic observations along with sequences of phonological features is defined in Section 2. This model includes contextual information aimed to constrain the search when identifying the articulator movement. These constraints have shown to be also useful for detection of audio events [14]. Section 3 includes a preliminary set of experiments for identification of articulatory features over a Spanish phonetic corpus. Experimental results show the importance of contextual information to detect phonological events. Moreover, they show that finite-state models can also be a promising framework to include phonological knowledge into ASR systems. Finally, Section 4 includes some concluding remarks and further work.

2. Hierarchical models to represent Phonological features

Under certain assumptions, it has been demonstrated that k-TSS models are equivalent to some extended n-gram models, where k stands for n. However, stochastic k-TSS regular languages as well as k-TSS SFSA benefit from the formal language theory. Thus, as mentioned above, efficient learning and decoding algorithms can be applied. In particular, composition between finite-state models is specifically interesting when combining several knowledge sources such as acoustic and phonological ones. Moreover, this composition can be carried out on-the-fly at decoding time in a very efficient manner [15]. In this framework we are now defining a hierarchical two-level stochastic finite state model which considers both phonological distinctive features and discrete observations derived from the mel-cepstrum space of representation.

Let $\bar{o} = o_1 o_2 \dots o_T$ be a sequence of T discrete acoustic observations associated to a spoken utterance. The a priori probability of $\bar{o}$ can be calculated as a product of conditional probabilities and then estimated according to a stochastic k-TSS model through eq. (1).

$$P(\bar{o}) = \prod_{i=1}^{T} P(o_i|o_1^{i-1}) \approx \prod_{i=1}^{T} P(o_i|o_{i-k_o+1}^{i-1}) \qquad (1)$$

where $h_i = o_1^{i-1} = o_1 \dots o_{i-1}$ represents the history of acoustic observations o_i and it is approached by the $k_o - 1$ predecessor observations. Let us now define a segmentation (s) of the sequence $\bar{o} = o_1 o_2 \dots o_T$ into N phrases or subsequences

of observations, as a vector of N indexes, $s = (s_1, \ldots, s_N)$, such that $s_1 \leq \ldots \leq s_N = T$. The sequence $\bar{o}$ can now be represented in terms of such segmentation as follows:

$$\bar{o} = o_1 \ldots o_T = o_{s_0=1}^{s_1} \ldots o_{s_{N-1}+1}^{s_N=T} \tag{2}$$

where $o_{s_{i-1}+1}^{s_i} = o_{s_{i-1}+1} \ldots o_{s_i}$ is a phrase. The set of all possible segmentations of a given sequence $\bar{o}$ is denoted as $S(\bar{o})$.

For simplicity, only phone-synchronous segmentations are considered in this work. As a consequence a phrase is just the subsequence of acoustic observations corresponding to a phoneme.

Let $C = \{C^1, \ldots, C^M\}$ be a set of previously defined classes $C^m = \{c_j^m\}$, where $m = 1, \ldots, M$ and $j = 1 \ldots |C^m|$. Each class $C^m \in C$ corresponds to a group of distinctive features according to the Phonology of the Language. Thus, each $c_j^m \in C^m$ represents a phonological feature of class C^m. As an example, if C^m is the class that represents sonority then c_j^m can take the values $voiced$ or $unvoiced$. Thus, $C^m = \{voiced, unvoiced\}$. Similarly if C^m stands for the place of articulation then c_j^m would take the values $bilabial$, $dental$, etc. and $C^m = \{bilabial, dental, velar, \ldots\}$.

Observations in the sequence $\bar{o} = o_1 o_2 \ldots o_T$ can be classified according to any of the phonological classes $C^m \in C$. Under a particular classification, a sequence of phonological features $\bar{c}^m = c_1^m c_2^m \ldots c_N^m$, associated to the sequence $\bar{o}$, is obtained. On the other hand, the sequence $\bar{o}$ can be associated to a sequence of phrases according to a segmentation as shown in eq. (2). Thus, in order to obtain a sequence of phonological features that matches the sequence $\bar{o}$ in eq. (2), a phrase or subsequence of acoustic observations $o_{s_{i-1}+1}^{s_i}$ need to be associated to a specific phonological feature $c_j^m \in C^m$. The set of all possible segmentations of a given sequence $\bar{o}$ compatible with the feature sequence $\bar{c}^m$ is denoted as $S_{\bar{c}^m}(\bar{o})$. Thus, we can associate a set of phrases or subsequences of acoustic observations to each $c_j^m \in C^m$.

When only phone-synchronous segmentations are considered, all of the sequences $\bar{c}^m$ share the same segmentation $s = (s_1, \ldots, s_N)$ and for any pair $(\bar{c}^m, s)$ associated to a sequence $\bar{o}$, N is just the number of phonemes in the utterance represented by the sequence $(\bar{o})$.

As an example consider the spanish word "cama" made up of the phonemes: /k/, /a/, /m/ and /a/. A specific segmentation and two different sequences of features associated to Sonority class and Place of articulation class, are given below:

/k/	/a/	/m/	/a/
unvoiced	*voiced*	*voiced*	*voiced*
velar	*vowel*	*bilabial*	*vowel*
$o_1 \ldots o_{s_1}$	$o_{s_1+1} \ldots o_{s_2}$	$o_{s_2+1} \ldots o_{s_3}$	$o_{s_3+1} \ldots o_{s4}$

where s_i are the indexes associated to the segmentation s. Thus, $o_1 \ldots o_{s_1}$ is a phrase associated to the phonological feature $unvoiced$ in class *Sonority* and being also associated to the phonological feature $velar$ in class *Place of articulation*.

Considering a specific set of features C^m the segmentation and the classification of a sequence of observations can be understood as hidden variables, as shown in [16, 12]. In this way, the probability of a sequence $\bar{o}$ can now be obtained by means of eq. (3):

$$\begin{aligned} P(\bar{o}) &= \sum_{\forall \bar{c}^m \in \mathcal{C}^{m*}} \sum_{\forall s \in \mathcal{S}_{\bar{c}^m}(\bar{o})} P(\bar{o}, \bar{c}^m, s) \\ &= \sum_{\forall \bar{c}^m \in \mathcal{C}^{m*}} \sum_{\forall s \in \mathcal{S}_{\bar{c}^m}(\bar{o})} P(\bar{o}, s|\bar{c}^m) P(\bar{c}^m) \\ &= \sum_{\forall \bar{c}^m \in \mathcal{C}^{m*}} \sum_{\forall s \in \mathcal{S}_{\bar{c}^m}(\bar{o})} P(\bar{o}|s, \bar{c}^m) P(s|\bar{c}^m) P(\bar{c}^m) \end{aligned} \tag{3}$$

being $\mathcal{C}^{m*}$ the set of all possible sequences of phonological features given a predetermined class C^m.

The term $P(s|\bar{c}^m)$ in eq. (3) could be estimated assuming that the segmentation probability is a positive constant (α) as considered in other works such as [17].

The probability of a given sequence of phonological features, $P(\bar{c}^m)$, can be calculated as a product of conditional probabilities. The probability of feature c_i^m given its history of features $c_1^m \ldots c_{i-1}^m = c_1^{i-1(m)}$ can be estimated through a Stochastic k-TSS model, where $(k_{c^m} - 1)$ stands for the maximum length of the sequence history considered as eq. (4) shows.

$$P(\bar{c}^m) = \prod_{i=1}^{N} P(c_i^m | c_1^{i-1(m)}) \simeq \prod_{i=1}^{N} P(c_i^m | c_{i-k_{c^m}+1}^{i-1(m)}) \tag{4}$$

This model considers the specific relations among phonological features in a language.

Finally, $P(\bar{o}|s, \bar{c}^m)$ is estimated in accordance with zero-order models. In this way, given a sequence of features $\bar{c}^m$ and a segmentation s, the probability of a subsequence of observations given a feature c_i^m, depends exclusively on feature c_i^m and not on the previous ones, as shown below.

$$P(\bar{o}|s, \bar{c}^m) \simeq \prod_{i=1}^{N} P(o_{s_{i-1}+1}^{s_i} | c_i^m) \tag{5}$$

To estimate this probability a stochastic k_{co^m}-TSS model can be used for each class, as shown in eq. (6),

$$P(o_{s_{i-1}+1}^{s_i} | c_i^m) \simeq \prod_{j=s_{i-1}+1}^{s_i} P(o_j | o_{j-k_{co^m}+1}^{j-1}, c_i^m) \tag{6}$$

where $(k_{co^m} - 1)$ stands for the maximum length of the acoustic observation history that is considered in each c_i^m. Let us note that the history is truncated to the feature boundaries.

Summing up, the probability of a sequence of acoustic observations can be computed by means of eq. (7):

$$P(\bar{o}) \simeq P_{M_{C^m}}(\bar{o}) = \alpha \sum_{\forall \bar{c}^m \in \mathcal{C}^{m*}} \sum_{\forall s \in \mathcal{S}_{\bar{c}^m}(\bar{o})} \prod_{i=1}^{N} \left[\left[\prod_{j=s_{i-1}+1}^{s_i} P(o_j | o_{j-k_{co^m}+1}^{j-1}, c_i^m) \right] P(c_i^m | c_{i-k_{c^m}+1}^{i-1(m)}) \right] \tag{7}$$

M_{C^m} model stands for the probability of sequences of acoustic observations given a particular feature class C^m being $m = 1 \ldots M$. Thus, M different models can be defined according to eq. 7.

This formulation assumes not only a clustering of acoustic features into the phonological space but also a certain relationship between sequences of observations that depends on each phonological feature. It allows a simple learning procedure of a set of small SFSA that could be composed with other SFSA. Then, on-the-fly composition of all the models is considered at

decoding time to incorporate phonological knowledge in ASR systems.

As a first application of this proposal we have formulated a set of phonological classifiers, i.e. detectors of articulatory features. The sequence $(\bar{o})$ can be considered to be produced by a Markov Chain, where its p_i states correspond to the SFSA associated with the stochastic model of choice, M_{C^m}. For each classification, i.e. phonological class C^m, and approaching sums in eq. (3) by a maximum, the specific segmentation and classification of a sequence of observations $(\bar{o})$ can be obtained through the Viterbi algorithm as eq. (8) shows:

$$[\hat{s}, \hat{\bar{c}}^m] = \arg \max_{\forall \bar{c}^m \in \mathcal{C}^{m*} \forall s \in \mathcal{S}_{\bar{c}^m}(\bar{o})} P(\bar{c}^m, s) P(\bar{o}|\bar{c}^m, s) = \quad (8)$$

where $P(\bar{o}|\bar{c}^m, s) = \sum_{p_0^T} P(\bar{o}, p_0^T|\bar{c}, s)$, T is the number of acoustic observations in $\bar{o}$ and $p_i^{i+j} = p_i, ..., p_{i+j}$ represents a sequence of p_i states which corresponds to the SFSA associated to M_{cm} model. Let us note that p_0^T will be different for different segmentations (s) and sequences of classes $(\bar{c})$.
Using the Viterbi algorithm the obtained $\hat{s}$ and $\hat{\bar{c}}^m$ are the segmentation and feature sequence that matches the best sequence of p_i states in the decoding network.

3. Experimental evaluation

The formulation proposed in the previous section was experimentally evaluated over a classification task of phonological features. For this purpose, the Albayzin [18] phonetic corpus was used. This is a phonetically balanced corpus consisting of utterances of the Castilian variety of Spanish. Table 1 contains a short description of the main characteristics of the corpus.

	Speakers	Sentences	Phonemes	Frames
Training	164	4800	187848	1465367
Test	40	2000	93696	711342

Table 1: Summary of the Albayzin phonetic corpus

Each utterance was sampled at 16 Khz and then parameterized to get sequences of mel-frequency cepstral coefficients with 25ms window and 10ms overlapping. The normalized energy as well as dynamic characteristics (first and second derivatives) were also computed resulting in a 39-component acoustic observation vector including cepstral coefficients, their first and second derivative, energy and its derivative. A standard and straightforward procedure of vector quantization was applied to get a discrete codebook. The associated codewords stands for the discrete acoustic alphabet used in this work.

For these experiments we considered five classes C^m, $m = 1 \ldots 5$ of distinctive features according to sonority, manner of articulation, place of articulation and specific front and open axis for vowels. The articulatory features and Spanish phonemes associated to each class are shown in Table 2.

Then we trained a k-TSS model for each feature and each class C^m. Thus, a set of $C^m \times |C^m|$ SFSA were obtained. Let us show a simplified example illustrated in Figure 1, where a SFSA considering the relations among the phonological features is shown. On the other hand, o_i symbols are codeword labels belonging to a finite alphabet $\Sigma_o = \{o_A, o_B, o_C, \ldots\}$ that stands for a codebook obtained from the procedure of vector quantization. According to C^1 classification in Table 2 two different SFSA associated to feature $c_1^1 = voiced$ and to feature $c_2^1 = unvoiced$ are defined. Focusing on c_1^1, note that it is made

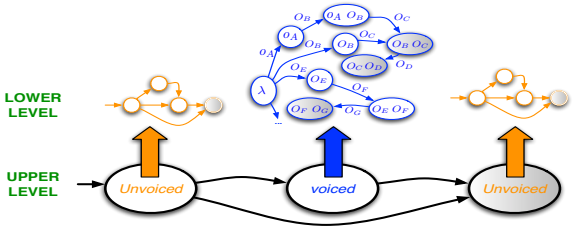

Figure 1: SFSA considering the relations among phonological features and the specific SFSA associated to each phonological feature $c_1^1 = voiced$ and $c_2^1 = unvoiced$ inferred from Σ^+ for $k_{co^m} = 3$. λ is the initial state and states in grey are final states

up of different phrases associated to voiced phonemes. Let us assume that, specifically for this distinctive feature, a training corpus $\Sigma^+ = \{o_A o_B o_C, o_A o_B o_C o_D, o_B o_C o_D\ o_E o_F o_G\}$ is employed consisting on subsequences of discrete observations associated to some voiced phonemes. Thus, a k-TSS SFSA can be inferred from Σ^+. Figure 1 shows this SFSA for $c_1^1 = voiced$ and $k_{co^m} = 3$.

$\mathcal{C}^1$ = *Sonority*	
c_1^1 = Voiced	a e i o u b d g l ll r rr m n ñ
c_2^1 = Unvoiced	p t k f z s j ch *sil*

$\mathcal{C}^2$ = *Vowel (Front axis)*	
c_1^2 = Front	e i
c_2^2 = Central	a
c_3^2 = Back	o u
c_4^2 = Consonantal	*rest*
c_5^2 = Silence	*sil*

$\mathcal{C}^3$ = *Vowel (Open axis)*	
c_1^3 = Open	e o
c_2^3 = Close	i u
c_3^3 = Midclose	a
c_4^3 = Consonantal	*rest*
c_5^3 = Silence	*sil*

$\mathcal{C}^4$ = *Manner*	
c_1^4 = Plosive	p t k b d g
c_2^4 = Fricative	f z s j
c_3^4 = Affricate	ch
c_4^4 = Lateral	l ll
c_5^4 = Trill	r
c_6^4 = M. Trill	rr
c_7^4 = Nasal	m n ñ
c_8^4 = Vowel	a e i o u
c_9^4 = Silence	*sil*

$\mathcal{C}^5$ = *Place*	
c_1^5 = Bilabial	p b m
c_2^5 = Labiodental	f
c_3^5 = Linguodental	z
c_4^5 = Alveolar	t d s ch l r rr n
c_5^5 = Palatal	ll ñ
c_6^5 = Velar	k g j
c_7^5 = Vowel	a e i o u
c_8^5 = Silence	*sil*

Table 2: Sets of classes, C^m, used in the experiments.

A first series of experiments were carried out using codebooks of different sizes. Figure 2*(a)* shows the frame classification accuracy obtained for each C^m class when we varied the size of the discrete observation alphabet and keep fixed both $k_{c^m} = 2$ and $k_{co^m} = 2$. This Figure shows that a size of 1024 codewords leads to the best frame classification accuracy for this task when any C^m class is considered.

We then set this value for the codebook size and carried out a second series of experiments considering different lengths for the history of phonological features, i.e. varying the value of the k_{c^m} in eq. (4) for each class C^m. Figure 2*(b)* shows the frame classification rates for each feature c_i^m. Different values of k_{c^m} were considered in these experiments while keeping $k_{co^m} = 2$. This Figure shows that given a C^m class similar accuracy rates were obtained for $k_{c^m} > 2$.

In a similar way different values of k_{co^m} in eq. (6) have also been considered for each phonological feature $c_j^m \in C^m$, $m = 1 \ldots M$ and $j = 1 \ldots |C^m|$. Figure 3*(a)* shows the frame classification accuracy obtained for class $\mathcal{C}^4$ (manner of articulation) and Figure 3*(b)* for class $\mathcal{C}^1$ (sonority), when keeping $k_{c^m} = 2$. Different values of the k_{co^m} parameters were considered for each c_i^m phonological feature. These Figures show that

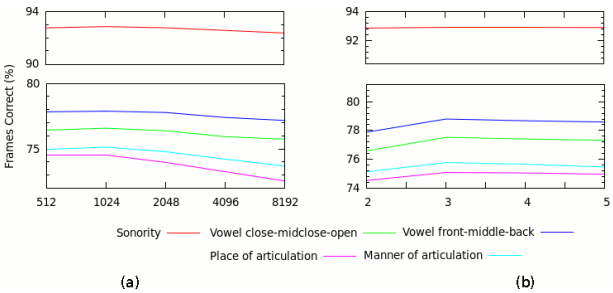

Figure 2: *(a)* Frame accuracy rates for C^m sets of classes varying the observation alphabet length, being $k_{c^m} = 2$ and $k_{co^m} = 2$ for all c_i^m; (b) Frame accuracy rates for each C^m varying the phonological history length k_{c^m}, being $k_{co^m} = 2$ for all c_i^m.

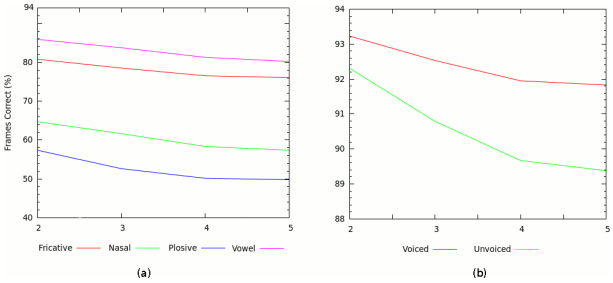

Figure 3: Frame accuracy rates for some distinctive features in class $\mathcal{C}^4$ (a) for $k_{c^4} = 2$, and in class $\mathcal{C}^1$ (b) for $k_{c^1} = 2$. Different lengths of observation histories were evaluated.

higher values of k do not lead to better accuracies. Note that all the experiments reported higher performance for the sonority class than for others. This corpus was previously used to classify phonological features trough dynamic neural networks [19]. Similar classification rates were reported for Sonority class but significantly higher rates were obtained for the other classes. This is probably due to the strong discretization carried out in this work by the procedure of vector quantization over vectors of 39 components.

4. Conclusions

We have formulated in this work a hierarchical model that considers both space of representations, the one based on phonological distinctive features and the one derived from the mel-cepstrum coefficients. The proposal is based on Stochastic finite-state models, specifically on k-TSS SFSA, which are considered to be a syntactic approach of n-gram models. This framework allows independent learning of the involved SFSA whereas they are dynamically combined on-the-fly at decoding time. From this formulation a classifier of articulatory features has been derived and then evaluated over a Spanish phonetic corpus. Despite of the strong discretization process, which can be solved in another way, we have obtained good classification rates for phonological features. Given the versatility and ability of combination of the finite-state models, this work shows a promising framework not only to detect phonological knowledge but also to include it into ASR systems.

5. Acknowledgements

This work has been partially supported by the Spanish MICINN under grants Consolider Ingenio CSD2007-00018 and TIN2008-06856-C05-01 and by the Basque Government under grants GIC10/158 IT375-10 and PIFA01/2008/034

6. References

[1] N. Chomsky and M. Halle, *The Sound Pattern of English*, Harper and Row, Eds., 1968.

[2] S. Stker *et al.*, "Integrating multilingual articulatory features into speech recognition," in *in Proc. Eurospeech*, 2003, pp. 1033–1036.

[3] R. Rose and P. Momayyez, "Integration of multiple feature sets for reducing ambiguity in asr," in *Proceedings of icassp*, vol. 4, Honolulu, USA, 2007, pp. 575–578.

[4] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer, Speech and Language*, vol. 14, pp. 333–353, 2000.

[5] J. Frankel *et al.*, "Articulatory feature recognition using dynamic bayesian networks," *Computer, Speech and Language*, vol. 21, pp. 620–640, 2007.

[6] J. Koreman and B. Andreeva, "Can we use the linguistic information in the signal?" *Phonus*, vol. 5, pp. 47–58, 2000.

[7] J. M. Olaso and M. I. Torres, "Integrating phonological knowledge in asr systems for spanish language," in *Progress in Pattern Recognition, Image Analysis, Computer Vision and Applications*, ser. Lecture Notes in Computer Science, 2010, vol. 6419, pp. 136–143.

[8] S. Siniscalchi and C.-H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for asr," *Speech Communication*, vol. 51, pp. 1139–1153, 2009.

[9] E. Vidal *et al.*, "Probabilistic finite-state machines - part II," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 27, no. 7, pp. 1025–1039, 2005.

[10] M. I. Torres and A. Varona, "k-tss language models in speech recognition systems," *Computer, Speech and Language*, vol. 15, no. 2, pp. 127–149, 2001.

[11] V. Guijarrubia and M. Torres, "Text and speech based phonotactic models for spoken language identification of basque and spanish," *Pattern Recognition Letters*, vol. 31, no. 6, pp. 523–532, 2010.

[12] R. Justo and M. Inés Torres, "Phrase classes in two-level language models for ASR," *Pattern Analysis and Applications*, vol. 12, no. 4, pp. 427–437, 2009, 10.1007/s10044-009-0165-y.

[13] A. Pérez *et al.*, "Joining linguistic and statistical methods for Spanish-to-Basque speech translation," *Speech Communication*, vol. 50, pp. 1021–1033, 2008.

[14] Q. Huang and S. Cox, "Using high-level information to detect key audio events in a tennis game," in *Proceedings of INTERSPEECH*, Makuhari, Japan, 2010.

[15] D. Caseiro and I. Trancoso, "Probabilistic finite-state machines - part II," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1281–1291, 2006.

[16] R. Justo and M. I. Torres, "Phrases in category-based language models for spanish and basque ASR," in *Proceedings of INTERSPEECH '07*, Antwerp, Belgium, August 2007, pp. 2377–2380.

[17] R. Zens and H. Ney, "Improvements in phrase-based statistical machine translation," in *Proceedings of HLT-NAACL'04*. Boston, MA: ACL, May 2004, pp. 257–264.

[18] F. Casacuberta *et al.*, "Desarrollo de corpus para investigación en tecnologías del habla (albayzin)." in *Procesamiento del lenguage natural*, vol. 12, 1992, pp. 35–42.

[19] J. M. Olaso and M. Inés Torres, "Speech production models for ASR in spanish language," in *Proceedings of VI Jorndas en Tecnologías del Habla and II Iberian SLTech*, Vigo, Spain, 2010, pp. 107–110.