# Hierarchical models based on a continuous acoustic space to identify phonological features

*Javier Mikel Olaso*[1], *María Inés Torres*[1]

[1]Departamento de Electricidad y Electrónica, Universidad del País Vasco, Spain

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

## Abstract

Phonological feature space has been proposed to represent acoustic models for automatic speech recognition (ASR) tasks. The most successful methods to detect articulatory gestures from the speech signal are based on Time Delay Neural Networks (TDNN). Stochastic Finite-State Automata have been effectively used in many speech-input natural language tasks. They are versatile models with well established learning algorithms that can easily be combined with other models. A two-level finite state model has also been proposed to classify articulatory features. However in this case a strong discretization procedure was required. In this work we propose a hierarchical finite-state model that considers two space of representations based on phonological features and on acoustic parameters, respectively. This model was evaluated in a phonological features identification task over a Spanish corpus. Experimental results show better frame classification accuracy than discrete models. Moreover, some specific articulations are better identified by the proposed models than by TDNN, leading to higher phone identification rates at frame level.

**Index Terms**: acoustic models, 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 movement 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] [4] [5] [6]. Moreover, the articulatory-based space seems to be more robust for noisy [7] or spontaneous speech and to be less variable in general [8] [6]. 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 methods to detect articulatory gestures from the speech signal are based on Time Delay Neural Networks (TDNN) [8]. Alternatively better recognition of articulatory features has been reported using dynamic Bayesian networks [9]. However, acoustic models based on phonetic features does not directly improve phonetic decoding performance. 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 [9] [3]. But significant improvements have been reported when combined with Hidden Markov models (HMM) defined over mel frequency cepstrum coefficients (MFCC) [3] [10]. Articulatory features have also been proposed with success for pronunciation variation models [5]. Phonoogical 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 [11].

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 [12], which is especially convenient for ASR [13] or speech translation tasks [14]. 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 [12] have been proposed for ASR [15], language identification [16], language modeling [13] or machine translation tasks [14].

In this context a two-level finite state model has also been proposed to classify articulatory features. This model was defined as a composition of $k$-TSS SFSA that considered discrete sequences of acoustic observations along with sequences of phonological features [17]. This model is summarized in Section 2.1. However in this case a strong discretization procedure was required. In this work we also propose a hierarchical finite-state model. The proposed model considers two space of representations based on phonological features and on acoustic parameters, respectively. The discrete phonological space is represented by a $k$-TSS model. A continuous acoustic observations space is now considered at the lower level, where a model is proposed for each phonological feature. As a consequence, the strong discretization required by models defined in [17] is no more needed. This proposal is presented in Section 2.2. Then Section 3 includes the results of a set of experiments for identification of articulatory features over a Spanish phonetic corpus. Finally, Section 4 deals with 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, 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. In this framework a hierarchical two-level stochastic finite state model has been defined [17], which considers both phonological distinctive fea-

tures and discrete observations derived from the mel-cepstrum space of representation. This proposal is summarized in Section 2.1. Then the continuous extension proposed in this paper is presented in Section 2.2

### 2.1. Hierarchical $k$-TSS models

Let $\bar{o} = o_1 o_2 \dots o_T$ be a sequence of T discrete acoustic observations associated to a spoken utterance. Also let $C = \{C^1, \dots, C^M\}$ be a set of previously defined classes $C^m = \{c_j^m\}$, where $m = 1, \dots, M$ and $j = 1 \dots |C^m|$. 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 $C^m$. Observations in the sequence $\bar{o} = o_1 o_2 \dots o_T$ can be classified according to any of the phonological classes $C^m \in C$. Under a particular classification, a sequence of phonological features $\bar{c}^m = c_1^m c_2^m \dots c_N^m$, associated to the sequence $\bar{o}$, is obtained. On the other hand, the sequence $\bar{o}$ can be associated to a sequence of $N$ phrases or subsequences of observations according to a segmentation $s = (s_1, \dots, s_N)$, such that $s_1 \leq \dots \leq s_N = T$.

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 [13]. In this way, the probability of a sequence $\bar{o}$ can now be obtained trough eq. (1):

$$P(\bar{o}) = \sum_{\forall \bar{c}^m \in \mathcal{C}^{m*}} \sum_{\forall s \in \mathcal{S}_{\bar{c}^m}(\bar{o})} P(\bar{o}|s, \bar{c}^m) P(s|\bar{c}^m) P(\bar{c}^m) \tag{1}$$

where $\mathcal{C}^{m*}$ is the set of all possible sequences of phonological features given a predetermined class $C^m$ and $\mathcal{S}_{\bar{c}^m}(\bar{o})$ is the set of all possible segmentations matching $\bar{c}^m(\bar{o})$.

The term $P(s|\bar{c}^m)$ in eq. (1) could be estimated assuming that the segmentation probability is a positive constant ($\alpha$) as considered in other works [18] [13].

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 $c_1^m \dots c_{i-1}^m = c_1^{i-1(m)}$ 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. (2) shows.

$$P(\bar{c}^m) = \prod_{i=1}^{N} P(c_i^m | c_1^{i-1(m)}) \simeq \prod_{i=1}^{N} P(c_i^m | c_{i-k_{c^m}+1}^{i-1(m)}) \tag{2}$$

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.

$$P(\bar{o}|s, \bar{c}^m) \simeq \prod_{i=1}^{N} P(o_{s_{i-1}+1}^{s_i} | c_i^m) \tag{3}$$

To estimate this probability a stochastic $k_{co^m}$-TSS model can be used for each class, as shown in eq. (4),

$$P(o_{s_{i-1}+1}^{s_i} | c_i^m) \simeq \prod_{j=s_{i-1}+1}^{s_i} P(o_j | o_{j-k_{co^m}+1}^{j-1}, c_i^m) \tag{4}$$

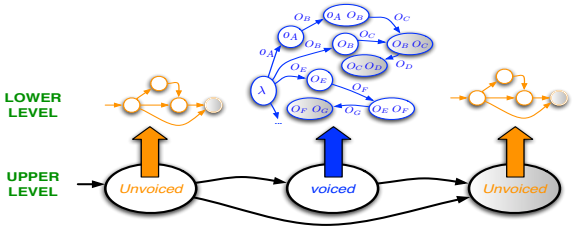


Figure 1: SFSA considering the relations among phonological features and the specific SFSA associated to each phonological feature $c_1^1 = voiced$ and $c_2^1 = unvoiced$ inferred from $\Sigma^+$ for $k_{co^m} = 3$. $\lambda$ is the initial state and states in grey are final states

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.

A $k$-TSS model for each feature and each class $C^m$ need to be trained and thus, a set of $m \times |C^m|$ SFSA obtained. Let us show a simplified example illustrated in Figure 1. The $o_i$ symbols are codeword labels belonging to a finite alphabet $\Sigma_o = \{o_A, o_B, o_C, \dots\}$ that stands for a codebook obtained from the procedure of vector quantization. Assume now that $C^1$ is the class that represents sonority , thus, $C^1 = \{voiced, unvoiced\}$. Then two different SFSA associated to feature $c_1^1 = voiced$ and to feature $c_2^1 = unvoiced$ need to be defined. Focusing on $c_1^1$, note that it is made 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 $\Sigma^+$. Figure 1 shows this SFSA for $c_1^1 = voiced$ and $k_{co^m} = 3$.

### 2.2. Continuous observations hierarchical models

In previous models, the discrete sequence $(\bar{o})$ can be considered to be produced by a Markov Chain, where its states correspond to the SFSA associated with the stochastic model of choice, $M_{C^m}$, as shown in Figure 1. However, a strong discretization process is required to represent the mel-cepstrum space as a discrete observations space. Alternatively a double stochastic process can be considered. Let now consider each state of the SFSA associated to a phonological as a generator of observed variables. In such a case the chain of states is not observable and another stochastic process associated to each state is producing the sequence of observations $(\bar{o})$. We have just transformed the low level SFSA into a Hidden Markov Model. We can now consider continuous observation densities, i.e. a Continuous Hidden Markov Model (CHMM).

Let $\mathcal{M}_{c_j^m}$ be a left-to-right CHMM defined for each phonological feature $c_j^m$ of class $C^m$. Such a model is defined by a set of states, a state transition probability distribution $A = a_{ij}$ where $a_{ij} = P[q_{t+1} = q_j | q_t = q_i]$ and an observation probability density function associated to each state represented as a finite mixture. Thus, the probability of emitting an observation vector $o_t \in (\bar{o})$ at a state $j$ is calculated according to

$$b(o_t)_j = \sum_{i=1}^{n} k_{ij} \mathcal{N}(o_t; \mu_{ij}, \Sigma_{ij}) \tag{5}$$

where $\mathcal{N}(o_t; \mu_{ij}, \Sigma_{ij})$ is the gaussian $i$ with mean vector $\mu_{ij}$ and covariance matrix $\Sigma_{ij}$.

Thus the subsequence $o_{s_{i-1}+1}^{s_i}$ is associated to one of the possible state sequence of the same length $\bar{q} = q_{s_{i-1}+1}^{s_i}$. Then $P(o_{s_{i-1}+1}^{s_i}|c_i^m)$ in Equation 3 is calculated as shown in Equation 6

$$P(o_{s_{i-1}+1}^{s_i}|c_i^m) \simeq P(o_{s_{i-1}+1}^{s_i}|\mathcal{M}_{c_j^m}) \tag{6}$$

where

$$\begin{aligned} P(o_{s_{i-1}+1}^{s_i}|\mathcal{M}_{c_j^m}) &= \sum_{\forall \bar{q}} P(o_{s_{i-1}+1}^{s_i}|\bar{q}, \mathcal{M}_{c_j^m}) \\ &= \sum_{\forall \bar{q}} \prod_{j=s_{i-1}+1}^{s_i} b(o_j) a_{q_j q_{j+1}} b(o_{j+1}) \end{aligned} \tag{7}$$

The topology of each CHMM defined for each phonological feature $c_j^m$ of class $C^m$ can be based on the structure of the $k$-TSS model inferred for $c_j^m$ when considering $(\bar{o})$ as a discrete sequence of observations (see Section 2.1) as shows Figure 1. However, the parameters of $\mathcal{M}_{c_j^m}$ can be underestimated for less frequent phonological features $c_j^m$. On the other hand a model $\mathcal{M}_{c_j^m}$ as simple as one state and a mixture of gaussians can also be proposed. But in such a case no temporal dependency in sequence $(\bar{o})$ is considered. In practice the number of states and the topology are part of the experimental layout which depends on the task goal.

Summing up, the probability of a sequence of acoustic observations can be computed by means of eq. (8):

$$P(\bar{o}) \simeq P_{M_{C^m}}(\bar{o}) = \alpha \sum_{\forall \bar{c}^m \in \mathcal{C}^{m*}} \sum_{\forall s \in \mathcal{S}_{\bar{c}^m}(\bar{o})} \prod_{i=1}^{N} \left[ \left[ \prod_{j=s_{i-1}+1}^{s_i} P(o_j|c_i^m) \right] P(c_i^m|c_{i-k_{c^m}+1}^{i-1(m)}) \right] \tag{8}$$

$M_{C^m}$ 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. 8.

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 and the acoustic space represented by the probability density function associated to each state. It allows a simple learning procedure of a set of CHMM that could be composed with other SFSA. Then, on-the-fly composition of all the models can be considered at decoding time to incorporate phonological knowledge in ASR systems in the same way as acoustic models are usually composed with Language Models in ASR systems based on SFSA [13].

As a first application of this proposal we have formulated a set of phonological classifiers, i.e. detectors of articulatory features. For each classification, i.e. phonological class $C^m$, and approaching sums in eq. (1) by the maximum, the specific segmentation and classification of a sequence of observations $(\bar{o})$ can be obtained through the Viterbi algorithm as eq. (9) shows:

$$\begin{aligned} [\hat{s}, \hat{\bar{c}}^m] &= \arg \max_{\forall \bar{c}^m \in \mathcal{C}^{m*} \forall s \in \mathcal{S}_{\bar{c}^m}(\bar{o})} P(\bar{c}^m, s) P(\bar{o}|\bar{c}^m, s) \\ &= \arg \max_{\forall \bar{c}^m \in \mathcal{C}^{m*} \forall s \in \mathcal{S}_{\bar{c}^m}(\bar{o})} P(\bar{c}^m, s) \sum_{\forall \bar{q}} P(\bar{o}, \bar{q}|\bar{c}, s) \end{aligned} \tag{9}$$

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 states $\bar{q}$ in the decoding network. Let us note that $\bar{q}$ will be different for different segmentations (s) and sequences of classes $(\bar{c})$.

# 3. Experimental results

The formulation proposed in the previous section was experimentally evaluated over a classification task of phonological features. For this purpose, the Albayzin [19] 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 25 ms 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. For these experiments we considered five classes $C^m$ 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.

| $\mathcal{C}^1$ = *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.

Two series of experiments where carried out to evaluated models defined in Section 2.2. In the first series we trained a $k$-TSS model (see Section 2.1) for each feature and each class $C^m$. Thus $m \times |C^m|$ SFSA (see Figure 1) representing discrete acoustic observations (DM) were obtained. Then five $k$-TSS where obtained to model the relations among phonological features for each of the five $C^m$ $m = 1 \ldots 5$ classes.

In the second series we trained $m \times |C^m|$ left-to-right CHMM (see Section 2.2) representing continuous acoustic observations (CM). In a previous tuning step we found that a combination of 32 gaussians leaded to the best frame classification accuracy for this task. Same five high level models for each of the five $C^m$ $m = 1 \ldots 5$ classes where also used in this series.

Table 3 shows frame classification accuracy obtained for the set of $|C^m|$ classes. For comparison purposes results previously obtained using Neural Networks (NN) classifiers were also included [17]. This table shows that higher phonological feature rates were obtained for all the $C^m$ classes when continuous acoustic observation were considered. However, these results did not outperform identification rates obtained trough classifiers based on neural networks.

| | CM | DM | NN |
|---|---|---|---|
| Sonority | 93.62 | 91.58 | 94.0 |
| V. central-front-back | 82.99 | 78.93 | 87.4 |
| V. open-close-midclose | 81.82 | 77.57 | 86.0 |
| Manner | 81.76 | 75.74 | 83.8 |
| Place | 82.14 | 75.17 | 83.4 |
| TOTAL | 84.47 | 79.80 | 86.92 |

Table 3: Frame classification accuracy for the set of $|C^m|$ classes. CM=Continuous Models, DM=Discrete Models, NN=Neural Networks.

Table 4 shows frame classification accuracy obtained for each of the $|C^m|$ classes. This table shows some significant differences between NN and CM behaviors. Even if NN classifiers obtained the best global accuracy CM models recognize better some *front axis* vowel features. Results for manner class shows that CM can identify significantly better *nasals* and the so difficult Spanish *trill*. Finally let us note that CM can classify less frequent features like *affricate*, *labiodental* or *linguodental* as well as the specific Spanish *palatal*. NN were not able at all of identify these features, maybe due, in some cases, to a low number of samples.

| *Sonority* | | | |
|---|---|---|---|
| | CM | DM | NN |
| Voiced | 93.12 | 91.50 | 92.0 |
| Unvoiced | 94.35 | 91.70 | 96.9 |

| *Vowel (Front axis)* | | | |
|---|---|---|---|
| | CM | DM | NN |
| Front | 84.77 | 77.16 | 81.1 |
| Central | 84.99 | 76.78 | 84.1 |
| Back | 80.82 | 62.52 | 73.8 |
| Consonantal | 77.75 | 79.40 | 90.8 |
| Silence | 95.32 | 91.03 | 96.0 |

| *Vowel (Open axis)* | | | |
|---|---|---|---|
| | CM | DM | NN |
| Open | 75.53 | 76.40 | 78.3 |
| Close | 83.15 | 60.47 | 75.1 |
| Midclose | 73.99 | 67.75 | 79.0 |
| Consonantal | 80.38 | 80.12 | 90.3 |
| Silence | 95.51 | 90.97 | 93.6 |

| *Manner* | | | |
|---|---|---|---|
| | CM | DM | NN |
| Plosive | 74.03 | 59.70 | 75.6 |
| Fricative | 90.93 | 82.66 | 92.5 |
| Affricate | 93.93 | 33.16 | 0.0 |
| Lateral | 72.14 | 33.87 | 54.0 |
| Trill | 66.56 | 17.88 | 47.0 |
| M. Trill | 64.37 | 9.48 | 0.0 |
| Nasal | 85.42 | 64.38 | 77.3 |
| Vowel | 77.74 | 86.14 | 89.7 |
| Silence | 94.57 | 91.91 | 95.7 |

| *Place* | | | |
|---|---|---|---|
| | CM | DM | NN |
| Bilabial | 69.56 | 39.62 | 73.2 |
| Labiodental | 65.47 | 34.74 | 0.0 |
| Linguodental | 87.38 | 54.32 | 0.0 |
| Alveolar | 77.23 | 68.03 | 83.0 |
| Palatal | 86.17 | 26.11 | 0.0 |
| Velar | 74.95 | 56.99 | 72.3 |
| Vowel | 82.97 | 85.73 | 91.0 |
| Silence | 95.88 | 92.47 | 96.8 |

Table 4: Frame classification accuracy for each of the $|C^m|$ classes. CM=Continuous Models, DM=Discrete Models, NN=Neural Networks.

Then we wanted to know the number of frames where all the features were correctly identified. In such a case the phoneme was also correctly classified. When at least one feature was incorrectly classified the phoneme was not recognized but the combination of detected features may represent another phoneme. In such a case an incorrect phoneme was detected. Finally the combination of detected features may not correspond to a phoneme, so that no phoneme is detected. Table 5 shows the phoneme identification accuracy at frame level. This table shows significantly better phoneme identification rates in terms of *all-features-correct* for the continuous models. The identification error has also a different distribution between *incorrect phoneme* identification and *no phoneme* identification for the models evaluated.

| | CM | DM | NN |
|---|---|---|---|
| Correct Phoneme | 64.91 | 58.61 | 58.65 |
| Incorrect Phoneme | 13.79 | 28.44 | 25.65 |
| No Phoneme | 21.3 | 12.95 | 15.71 |

Table 5: Phoneme identification accuracy at frame level. The phoneme was correct when all detected features were correct. It was incorrect when the detected features represent another phoneme. No phoneme was detected when detected features did not correspond to a phoneme. CM=Continuous Models, DM=Discrete Models, NN=NeuralNetworks.

# 4. Conclusions

Stochastic Finite State Automata have been successfully used when dealing with speech processing tasks. In this context a two-level finite state model had also been proposed to classify articulatory features. This model was defined as a composition of k-TSS SFSA. In this work we have reformulate these models to consider two space of representations based on phonological features and on acoustic parameters, respectively. The high level model representing the discrete phonological space is based on Stochastic Finite-State models, specifically on $k$-TSSS SFSA, which are considered to be a syntactic approach of $n$-gram models. A continuous acoustic observations space has now been considered at the lower level, where a model has been proposed for each phonological feature. This framework allows independent learning of the involved models whereas they can be 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. The experiments carried out showed good identification rates for phonological features, mainly when continuous acoustic observations were considered. The proposed models did not improve global frame classification accuracy obtained trough classifiers based on neural networks. However, they are able to identify difficult places and manners of articulation that neural networks can not identify, thus leading to significantly higher phone identification rates at frame level. 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's.

# 5. Acknowledgements

This work has been partially supported by the Spanish MICINN under grant TIN2011-28169-C05-04 and by the Basque Government under grant IT685-13.

# 6. References

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

[2] S. Stker, F. Metze, T. Schultz, and A. Waibel, "Integrating multi-

lingual 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] A. Naess, K. Livescu, and R. Prabhavalkar, "Articulatory feature classification using nearest neighbors," in *Proceedings of INTERSPEECH*. ISCA, 2011, pp. 2301–2304.

[5] P. Jyothi, K. Livescu, and E. Fosler-Lussier, "Lexical access experiments with context-dependent articulatory feature-based models," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[6] D. Yu, S. M. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting attribute and phone estimation accuracy with deep neural networks for detection-based speech recognition," in *Proceedings of icassp*, 2012.

[7] V. Mitra, H. Nam, C. Y. E. Wilson, E. Saltzman, and L. Goldstein, "Articulatory information for noise robust speech recognition," *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 7, pp. 1913–1924, 2011.

[8] 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.

[9] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic bayesian networks," *Computer, Speech and Language*, vol. 21, pp. 620–640, 2007.

[10] P. K. Ghosha and S. Narayanan, "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *Journal of the Acoustical Society of America*, vol. 130 (4), p. EL257, 2011.

[11] 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.

[12] E. Vidal, F. Thollard, C. de la Higuera, F. Casacuberta, and R. Carrasco, "Probabilistic finite-state machines - part II," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 27, no. 7, pp. 1025–1039, 2005.

[13] 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.

[14] A. Pérez, M. Torres, and F. Casacuberta, "Joining linguistic and statistical methods for Spanish-to-Basque speech translation," *Speech Communication*, vol. 50, pp. 1021–1033, 2008.

[15] 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.

[16] 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.

[17] J. M. Olaso, M. I. Torres, and R. Justo, "Representing phonological features through a two-level finite state model." in *Proceedings of INTERSPEECH*. ISCA, 2011, pp. 1733–1736.

[18] 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.

[19] F. Casacuberta, R. Garcia, J. Llisterri, C. Nadeu, J. Pardo, and A. Rubio, "Desarrollo de corpus para investigación en tecnologías del habla (albayzin)." in *Procesamiento del lenguage natural*, vol. 12, 1992, pp. 35–42.