

## ACOUSTIC MODELS OF CV SYLLABLES FOR SPEECH RECOGNITION

S. Fernández<sup>1</sup>, S. Feijóo<sup>2</sup>, and N. Barros<sup>2</sup>

<sup>1</sup>IDSIA

Dalle Molle Institute for Artificial Intelligence  
Galleria 2, 6928 Manno, Switzerland  
e-mail: santiago@idsia.ch

<sup>2</sup>Department of Applied Physics

University of Santiago de Compostela  
15782 Santiago de Compostela, Spain  
e-mail: fasergio@usc.es

**Keywords:** Speech recognition, acoustic models, coarticulation.

**Abstract.** *Auditory identification of phonemes depends on the target phoneme and on the signal surrounding it. Two sources of information have been proposed to explain this fact: a) coarticulatory information in adjacent segments; b) the phonetic identity of the segments. It is not clear yet if the target and adjacent segments are processed jointly, or if they are processed independently and the corresponding information is later combined. Speech recognition systems based on context-dependent phonetic models opt for the later alternative, and the information is combined assuming that the target and adjacent segments are independent. We propose models of interaction that take into account both the coarticulatory information and the phonetic identity of the segments, processed either independently or jointly. The acoustic characterization of the segments was based on the coding of the spectral energy. The models were first tested in fricative-vowel CV syllables, showing that the best automatic recognition rates were achieved when both sources of information were included and the C and V segments were processed jointly. Later, this model was tested with a database of continuous speech and compared with the performance of a context-dependent HMM-based system: 25% and 17% reduction in error rate was achieved for clean speech and for speech in noise, respectively.*

### 1 INTRODUCTION

Automatic recognition of phonemes in continuous speech is a complex problem due to the absence of a one-to-one mapping between acoustic segments and phonetic segments. It is well known that the auditory identification of phonemes depends not only on the target phoneme but also on adjacent segments, an effect that we shall call phonetic integration. Two sources of information have been proposed to explain this effect: a) coarticulatory information in adjacent segments; and b) the phonetic identity of adjacent segments<sup>[1,2]</sup>. In a previous work<sup>[3]</sup>, listeners benefited from the presentation of the vocalic segment to identify the fricative in consonant-vowel (CV) syllables. Hidden Markov models (HMM) were implemented for the two types of stimuli (C and CV). In the CV condition the model included the coarticulatory information of the vocalic segment, processing the CV segment as a whole. Results for the different acoustic models showed a similar trend to listeners' perception of the segments, with higher recognition rates in the CV condition. The proposed model was similar to that of Nossair & Zahorian for stop consonants<sup>[4]</sup>, although their classifier, based on quadratic discriminant analysis, was superior to the one used in the HMM-based system, and their approach was not based on a finite number of states to describe the speech dynamics but on the trajectory of the acoustic parameters along the CV segment.

In this paper, the previous work is augmented with the inclusion of the phonetic identity of the vowel segment as a source of information on fricative identity. Acoustic models that process the C and V segments separately and later combine both sources of information were developed and compared with acoustic models that process the CV segment as a whole. Two methods for representing speech dynamics were considered: HMMs, and the method proposed by Nossair & Zahorian.

Nevertheless, the following question arises: how can phonetic integration be implemented in Automatic Speech Recognition (ASR) systems? A first approach in ASR systems consists in using mono-phone models, where each model represents a particular phonetic category. This approach does not take into account the influence of adjacent segments of speech on the target phoneme. A method for including the dynamic characteristics of speech production is to consider different models of phonetic units for every context, indicating that the target phonetic segment has been uttered in the context of (for tri-phone models) a previous and next phoneme. Since the dynamics of speech production entails that there is acoustic information not only on the segment corresponding to a given phoneme, different alternatives have been proposed to model the trajectory of the acoustic parameters, including the transitional regions between phonemes<sup>[5]</sup>, or considering temporal intervals that include the whole sentence<sup>[6]</sup>. Several of these models use a previous segmentation of the speech signal that can be obtained, for instance, with tri-phone models. Nevertheless, the segmentation obtained with tri-phone models is far from optimum<sup>[7]</sup>, and significant improvements in the classification rate, as compared to that of a baseline tri-phone system, rely on the use of a reference segmentation of the speech signal obtained by forced alignment or manually.

An important drawback of the tri-phone and mono-phone systems is that they assume a statistical independence between phonetic segments, whereas it is clear that there is a statistical dependence between the phonemes and their contexts<sup>[2]</sup>. A possible alternative is to use syllables as the recognition units, but the high number of training units required makes this approach problematic<sup>[8]</sup>. Hybrid systems using both syllables and phonemes have also been used with better results than those obtained by the tri-phone systems.

The model proposed in this paper uses phonemes as recognition units to keep a reduced set of categories. It is based on the inclusion of the relevant information present in segments adjacent to that of the target phoneme. In contrast with tri-phone models, the assumption of independence between segments at the phonetic level is relaxed. The model is integrated into the framework used by current ASR systems.

## 2 EXPERIMENTS WITH FRICATIVE SYLLABLES

### 2.1 Materials and method

Stimuli consisted in fricative-vowel syllables (FV) formed by the combination of /θ, f, s, ʃ, x/ with /a, e, i, o, u/. The training set was formed by the repeated utterances of 29 male and 27 female speakers (2800 stimuli), and the test set was formed by the utterances of 10 male and 10 female speakers (500 stimuli) who were not part of the training set. Signals were sampled at 32 kHz and normalized to prevent differences in amplitude among samples. The segments corresponding to the fricative and the vowel were manually annotated in all the stimuli.

Twenty one listeners performed perceptual experiments with the test set in two conditions of presentation: a) the C condition formed by segments consisting only in the fricative noise; and b) the CV condition, in which stimuli consisted in the fricative noise plus 100 ms of the accompanying vowel. They were asked to respond according to the fricative they perceived in the stimuli.

To obtain the acoustic parameters the signal was down-sampled to 20 kHz, and 27 Mel-scaled cepstrum coefficients (MFCC) were computed on windows 15 ms long, overlapped 10 ms. The HTK software package (<http://htk.eng.cam.ac.uk>) was used for the HMM analysis, using models with 4 and 8 Gaussians. The probabilities of observation

were modeled with diagonal covariance matrices. Left-to-right models with three states were used for the C segments, two states for the V segments, and five states for the CV segments.

The objective of the system is to identify the fricative in a fricative-vowel syllable. Different models were considered:

*Model C*: only the acoustic information of the fricative noise is used; the classification groups are the 5 fricatives; no explicit information about the identity of the vowel is included; coarticulatory information is not included.

*Model Cq*: only the acoustic information of the fricative noise is used; the classification groups are the 25 fricative vowel syllables; explicit information about the identity of the vowel is included; coarticulatory information is not included. The probability of membership of a particular segment into one of the five fricative categories is computed adding the probabilities assigned by the model to all classification groups including that particular fricative.

*Model CV*: The acoustic information in both the fricative and the vowel is used jointly; the classification groups are the 5 fricatives; no explicit information about the identity of the vowel is included; coarticulatory information is included.

*Model CVq*: The acoustic information in both the fricative and the vowel is used jointly; the classification groups are the 25 fricative-vowel syllables; explicit information about the identity of the vowel is included; coarticulatory information is included.

*Model C-V*: The acoustic information in the fricative and in the vowel is obtained independently for each segment and later combined using a AND function, which means that the *a posteriori* probability of identifying the fricative,  $P$  is obtained as  $P_{CV} = P_C \cdot P_V$ , where  $P_C$  is the *a posteriori* probability of identifying the fricative from the information contained in the fricative, and  $P_V$  is the *a posteriori* probability of identifying the fricative from the information contained in the vowel; the classification groups are the 5 fricatives; no explicit information about the identity of the vowel is included; coarticulatory information is included.

*Model C-Vq*: The same as the model C-V, but considering the 25 fricative-vowel syllables as the classification groups; explicit information about the identity of the vowel is included; coarticulatory information is included.

## 2.2 Results

Table 1 shows the results obtained in the classification of fricatives by both the listeners and the automatic methods based on HMM. Also shown is the correlation between listeners' responses and the automatic recognition. Results using Nossair & Zahorian's method were slightly worse and are not included in the table. For model C, the classification rate was 72.0% and the correlation coefficient was 0.67; for model CV, 82.8% and 0.79, respectively; and for model C-V, the classification rate was 78.8% and the correlation coefficient was 0.78. In this case, Nossair & Zahorian's approach consisting in describing speech dynamics as the trajectory of the acoustic parameters did not obtain better results than using a HMM with a finite number of states.

Model	HMM (%)	Listeners (%)	Correlation (r) HMM vs. Listeners
C	80.4	63.3	0.77
Cq	81.8		0.76
CV	81.2	86.0	0.81
CVq	83.0		0.84
C-V	79.4		0.80
C-Vq	80.4		0.80

Table 1: Percent correct recognition obtained by both listeners and the HMM in the classification of fricatives

Listeners have a special difficulty in identifying the fricative from only the fricative noise (63% correct rate), and they benefit particularly from the inclusion of the vowel (86% correct rate). Curiously enough, the automatic method outperforms listeners in the C condition (80.4%, 81.8% correct rate). Nevertheless, when both sources of information, C and V, are included, the overall results of the automatic methods lie below those of the listeners. The best result is obtained by the model CVq (83% correct rate), which includes explicit information about the identity of the vowel, coarticulatory information, and both the C and V segments are evaluated jointly. This model also achieves the best correlation with listeners' responses (0.84). If no information about vowel identity is used, the results are slightly worse (model CV, 81.2% and  $r=0.81$ ). The lowest rates correspond to the models that treat the C and V segments separately (C-V and C-Vq).

Therefore, the automatic methods benefit from the inclusion of the two sources of information present in the V segment, namely coarticulatory and vowel identity information. Moreover, the information present in the C and V segments should be jointly treated in an integrated way to obtain the best results. The correlation obtained between listeners' responses and the automatic method indicates until which point the computational method mimics listeners' perception: the best results are obtained in the same conditions as above. Is that correlation satisfactory enough? One way to find out is to compare the correlation coefficient with the inter-listener consistency. Inter-listener consistency is obtained by dividing the sample of the listeners that participated in the perceptual experiment in two halves and calculating the corresponding correlation coefficient between the responses of the two groups. The correlation obtained was 0.91 in the C condition, and 0.97 in the CV condition. It is clear that, although the classification rates of the best automatic methods are close to those of listeners, they are not able to fully model the perceptual mechanisms that listeners use.

### 3 EXPERIMENTS WITH AN HMM RECOGNIZER ON CONTINUOUS SPEECH

From the results of the previous experiment it is clear that, in order to take advantage of the phonetic integration process, both the coarticulatory information and the information about vowel identity should be jointly taken into account when designing the acoustic models for the recognizer.

#### 3.1 Acoustic model

How can the information about vowel identity and the coarticulatory information be integrated in the model? To use syllables as recognition units is not a practical alternative, as was already discussed in the introduction. In order to keep a reasonable number of training units, the phoneme should be used. It should be noted that listeners do not need to hear the whole adjacent segment to use the contextual information, which means that they won't possibly be able to determine its precise identity in some instances, but they are able to extract at least part of its relevant characteristics. It is generally agreed that coarticulatory information in a vocalic segment due to a previous consonant should correspond to the transitional (initial) part of the vowel, while vowel identity is more or less present in the whole vocalic segment. Moreover, the effect of the following vowel on fricative recognition by listeners is not constant for different vowels, indicating that the interaction that takes place between the consonant and the vowel is quite complex. Since there is no way to know which acoustic characteristics of a vowel correspond to the coarticulatory information, and which particular characteristics correspond to vowel identity, the only practical way is to include both types of information as encoded by the acoustic parameters extracted from the signal.

The proposed model is based on the inclusion of part of the signal corresponding to the previous and to the next segments adjacent to the target phoneme, in such a way that both the characteristics related to the dynamics of the acoustic parameters, and certain characteristics related to the identity of the adjacent segments, might be modeled together with the characteristics of the target phoneme. The models are similar to tri-

phone models, except for the fact that now a given sentence is represented by a sequence of phonetic models that are associated to overlapping segments of the signal (see figure 1). This fact makes the automatic segmentation of the signal difficult. A possible solution is to use the segmentation obtained through tri-phone models and then selecting each phonetic segment plus half of the preceding and posterior segments. This new segment will be compared with the proposed models to determine the target phoneme. The rest of the decoding process is similar to that used with tri-phone models. The new model will be called *extended tri-phone* or *extended model*.

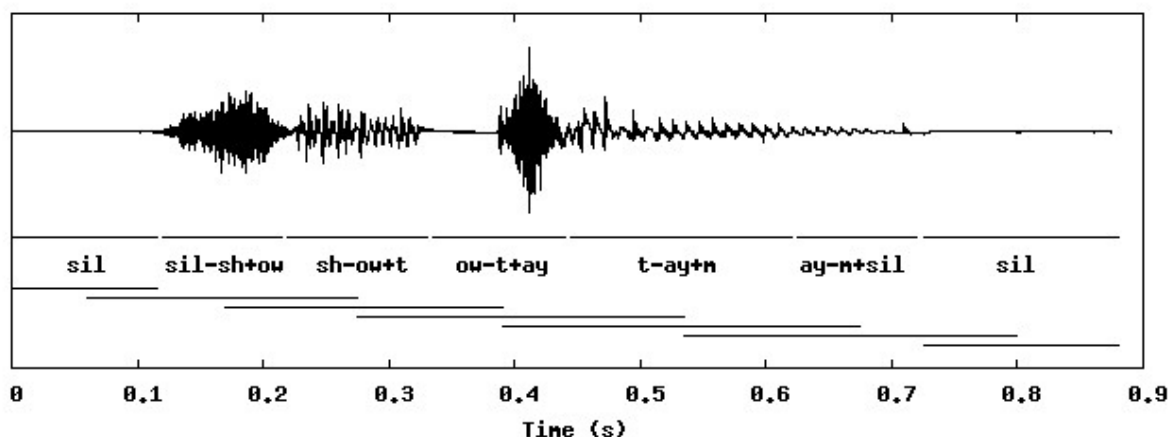


Figure 1. Comparison of speech segments modeled by tri-phones (top) and extended tri-phones (bottom) in the word “show time”.

### 3.2 Materials and method

The acoustic signals used in the experiments were obtained from the English version of the Wall Street Journal database (WSJCAM0). The selected recordings were done with a close-talking microphone. The training set was formed by 7861 sentences spoken by 92 talkers. The test set was formed by 368 sentences spoken by 20 talkers (ref: SI\_DT5A). The database includes the phonetic segmentation of the sentences obtained by forced alignment. The phonetic inventory includes 44 phonetic categories plus the “silence” category.

The models were trained with the HTK package (<http://htk.eng.cam.ac.uk/>). 12 MFCC coefficients plus the energy and the  $\Delta$  and  $\Delta\Delta$  coefficients of the MFCC and energy parameters were computed in 25 ms-long windows, with a 15 ms overlap between consecutive windows. First of all, a set was formed with 44 HMM mono-phone models (left-to-right, 3 states) plus one model for the “silence” category. Probabilities of observation were modeled by only one Gaussian using a diagonal covariance matrix. Tri-phone left-to-right models with 3 states were created from the original mono-phone models, with probabilities of observation modeled by only one Gaussian. The grouping technique was based on decision trees in order to link between them similar states and to have a robust estimate of the models. The decision tree is similar to that included as an example in the HTK package, but adapted to the phonetic categories included in our phonetic inventory. 17287 tri-phone models were obtained to represent all of the 85185 tri-phone models that may be formed with 45 phonetic categories (including “silence”). Finally, the number of Gaussians was increased gradually to eight.

The implementation of the extended tri-phone models was carried out similarly but while the tri-phone models are obtained with an embedded training process, the extended tri-phones were obtained with isolated training. The acoustic segments modeled by the extended tri-phones are obtained from the training sentences already segmented. The extended models consist of 7 states to represent the initial, middle and final states of the central phonetic segment, plus the transitional and stable regions of the signal

corresponding to the two adjacent phonetic segments. After grouping, the total number of models is 36585 to represent 85185 models. In this case the grouping of states implies that, for a given model, the identity of the adjacent segments may not be unequivocally determined.

Decoding of the test signals with the tri-phone models is carried out using the implementation of the Viterbi algorithm called *token passing*. For the extended models, decoding starts with a previous segmentation that determines the signal segments that include part of the adjacent phonemes. Decoding of that succession of segments is carried out with a specific decoder using the Viterbi algorithm.

The recognition carried out was purely phonetic, without employing any grammatical or other type of additional information different from the observation probabilities of the phonetic units.

### 3.3 Results

The correct recognition rate using the tri-phone models (considering insertions, substitutions and deletions) was 71.86%. The segmentation obtained with the tri-phones was used to decode the sequence with the extended tri-phones. The correct recognition rate with the extended models was 71.33%, very similar to that of the tri-phones. No significant differences were found between the recognition rates for each phonetic category obtained by the two models. The fact that the new models could not improve the results of the tri-phones may be due to the segmentation obtained, which is optimum for tri-phones but not for other models. One way of having a suitable segmentation for the extended models is to extract different alternative segmentations with tri-phones, and re-evaluate them with the extended models. A faster alternative that provides an upper bound of the improvement in classification rate that can be obtained with the proposed models, is to use the segmentation included along with the database and that was obtained by forced alignment (i.e. this segmentation was obtained automatically by forcing the system to recognize the labeled utterances).

The correct recognition rate using tri-phones with the new segmentation was 71.99%, only slightly better than the result obtained when the recognition process also has to determine the best segmentation. For the extended tri-phones, the recognition rate was 78.85%, higher than the previous result. It is clear that tri-phone-based systems are able to find a segmentation that maximizes the recognition rate when tri-phone models are used, although this segmentation may not be the most adequate for alternative models or from an acoustic-phonetic point of view. The extended models, though, achieve an important reduction of 25% in their error rate when a more appropriate segmentation is used. Figure 2 shows the differences in recognition rate of the two models for each phonetic category.

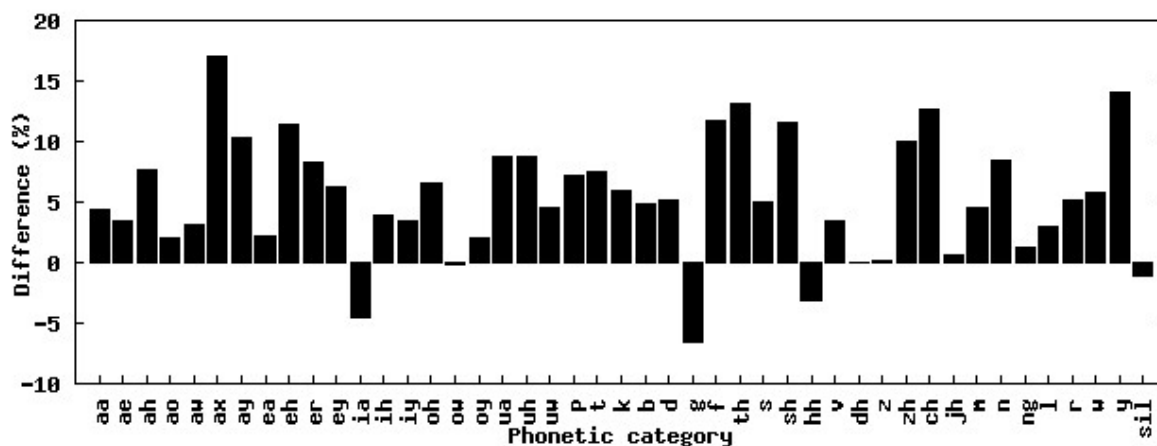


Figure 2. Differences per phonetic category between extended tri-phone models and tri-phone models.

Positive differences in the figure indicate that the extended models outperform the tri-phone models. There is a general improvement in all phonetic categories, except for /ia/, /g/, /hh/, and the "silence" category. Improvements higher than 10% are obtained for /ax/, /eh/, /f/, /th/, /sh/, /ch/, and /y/ (/th/ and /f/ are one of the most confusing pair of phonemes). Little improvement, if any, was found for /ow/, /dh/, /z/ and /jh/. The number of correctly recognized phonemes by each model is significantly different ( $\chi^2=96.48$ ,  $df=44$ ,  $p<8.5\times 10^{-6}$ ). There is no clear pattern, though, of which phonemes benefited most from the inclusion of the adjacent segments in the model. For vowels, there is an overall significant improvement ( $\chi^2=53.47$ ,  $df=19$ ,  $p<4.0\times 10^{-5}$ ), while for consonants there is no significant difference between the two methods ( $\chi^2=12.36$ ,  $gl=23$ ,  $p<0.96$ ), although it is clear that the confusing pairs /th/-/f/, /ch/-/sh/ and /s/-/sh/ benefit from the new models. Torre-Toledano *et al.*<sup>[13]</sup> concluded that the segmentation obtained by tri-phones usually model either the target phonetic segment plus part of the adjacent segment, or only part of the target segment (note that tri-phone models correspond to consecutive and non-overlapped segments of the acoustic signal). The inclusion of part of the adjacent segments takes place for the less stationary and more dynamic phonemes (stops, affricates, etc.). This was a systematic effect, indicating the importance of using the context for phonetic recognition. But, if the initial (transitional) part of the vowel in CV segments is assigned to the C segment, those vowels will be modeled without its initial part. When the extended tri-phone models are used, since consecutive segments overlap, there is a general improvement in the classification rate of vowels due to the possibility of including the whole vocalic segment (along with information from adjacent segments).

Taking into account that the use of context might be specially important for the robustness of the recognition system before noise, further experiments were carried out. The models trained on clean speech were used to classify noisy speech. White noise was added to the test set to achieve a signal-to-noise ratio of 20 dB. The reference system using tri-phones obtained a correct classification rate of 42.96%. Using the segmentation obtained with tri-phones, the extended models obtained a recognition rate of 45.46%. In this case, the extended models achieve better results with the segmentation obtained with tri-phone models.

In a second step, the segmentation obtained through forced recognition was used. Tri-phone models obtained a correct recognition rate of 47.71%, improving the results obtained with its own segmentation, which confirms the difficulties of tri-phones to obtain a suitable segmentation in the presence of noise. When the extended models were used, the correct recognition rate raised to 56.90%: an error reduction rate of 17% is achieved with respect to tri-phones. With an adequate segmentation, the new models are more robust, probably because they can take advantage of the information contained in parts of signal that are less affected by noise. For instance, stop burst are sometimes of little energy and short duration, which makes them prone to be masked by external noise. In this case the stop information contained in the transitional region due to coarticulation may act as an aid in recognition (see figure 3).



recognition systems. One of the weak points of current speech recognition systems is that they assume statistical independence between consecutive phonetic segments, whereas there is strong evidence for the opposite. In this framework, it is difficult to include the information obtained in acoustic-phonetics experiments into a totally different approach, such as that of HMM. It is clear that one of the strong points of the HMM models is their ability to encode the sequence of events that take place in a spoken sentence, but this ability is based on certain constraints that do not fit well with the facts of phonetic perception.

An alternative method to model dependencies in the speech signal is the use of recurrent neural networks (RNN) that can integrate information “dynamically”. In practice, RNNs have difficulties to use information from events occurring more than a few tens of time steps in the past. The LSTM-RNN<sup>[10]</sup> has been successfully applied to a number of problems requiring longer memory. Moreover, it outperforms RNNs in framewise phoneme classification<sup>[11]</sup>. This is a promising alternative methodology to model dependencies among phonetic segments and other long-term dependencies in speech signal processing and language processing.

## 5 ACKNOWLEDGMENTS

Part of this work was carried out when S. Fernandez was at the Department of Phonetics & Linguistics at University College London funded by the Marie Curie program of the EU: “Improving the human research potential and the socio-economic knowledge base”, under contract number HPMF-CT-2002-02129.

This work was partially funded by SNF grant 200020-100249 to J. Schmidhuber.

## REFERENCES

- [1] Nearey, T.M. (1992), “Context effects in a double-weak theory of speech perception”, *Language and Speech*, Vol. 35, pp. 153-171.
- [2] van Son, R.J.J.H., Pols, L.C.W. (1999), “Perisegmental speech improves consonant and vowel identification”, *Speech Communication*, Vol. 29, pp. 1-22.
- [3] Fernandez, S., Feijoo, S. (2003), “Comparing HMM-based recognition to perceptual phonetic integration in fricative-vowel syllables”, *Proceedings of the 15<sup>th</sup> International Congress of Phonetic Sciences, Barcelona (Spain)*, pp. 1433-1436.
- [4] Nossair, Z.B., Zahorian, S.A. (1991), “Dynamic spectral shape features as acoustic correlates for initial stop consonants”, *Journal of the Acoustical Society of America*, Vol. 89, pp. 2978-2991.
- [5] Reinhard, K., Niranjana, M. (2002), “Diphone subspace mixture trajectory models for HMM complementation”, *Speech Communication*, Vol. 38, pp. 237-265.
- [6] Deng, L., Ma, J. (2000), “Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics”, *Journal of the Acoustical Society of America*, Vol. 108, pp. 3036-3048.
- [7] Torre Toledano, D., Hernandez Gomez, L.A., Villarubia Grande, L. (2003), “Automatic phonetic segmentation”, *IEEE Transactions on Speech and Audio Processing*, Vol. 11, pp. 617-625.
- [8] Ganapathiraju, A., Hamaker, J., Picone, J., Ordowski, M., Doddington, G.R. (2001), “Syllable-based large vocabulary continuous speech recognition”, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, pp. 358-366.
- [9] Smits, R. (2001), “Hierarchical categorization of coarticulated phonemes: a theoretical analysis”, *Perception & Psychophysics*, Vol. 63, pp. 1109-1139.
- [10] Hochreiter, S. and Schmidhuber, J. (1997), “Long Short-Term Memory”, *Neural Computation*, Vol. 9(8), pp. 1735-1780.
- [11] Graves, A. and Schmidhuber, J. (2005), “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”, *Neural Networks* (in press).