# Syllable classification using static matrices and prosodic features

*Bogdan Ludusan, Antonio Origlia, Francesco Cutugno*

LUSI-lab, Department of Physical Sciences, "Federico II" University, Naples, Italy
{ludusan, cutugno}@na.infn.it, antori@gmail.com

## Abstract

In this paper we explore the usefulness of prosodic features for syllable classification. In order to do this, we represent the syllable as a static analysis unit such that its acoustic-temporal dynamics could be merged into a set of features that the SVM classifier will consider as a whole. In the first part of our experiment we used MFCC as features for classification, obtaining a maximum accuracy of 86.66%. The second part of our study tests whether the prosodic information is complementary to the cepstral information for syllable classification. The results obtained show that combining the two types of information does improve the classification, but further analysis is necessary for a more successful combination of the two types of features.

**Index Terms**: syllable, SVM, prosodic features

## 1. Introduction

Automatic Speech Recognition needs to work with units which have at the same time an intrinsic acoustic coherence and a significant correspondence with small linguistic units (usually phones) belonging to a specific language. The speech signal is usually partitioned into short, sub-phonic, segments and the recognition process attempts to assign a label, having a certain probability, to each of these segments. A bottom-up procedure (usually Viterbi search or a similar algorithm) combines these labels to form larger linguistic units, while a top-down procedure (a Language Model) provides further constraints to refine the decision process. In this view only two parallel processes, and, between these, only one depending on time, are combined to determine the output sentence.

This 'traditional' approach presents some limitations as it does not exploit linguistic information spread across the different time scales present within the speech signal. Studies regarding Human Speech Recognition [1][2] clearly indicate that the brain does not fall into the same error: various cortical areas, even in different brain hemispheres, process the signal at the same time with different temporal resolution. This led us to investigate ways in which we could extract information from the speech signal using a temporal resolution corresponding to a linguistic unit longer than the phone. We chose the syllable, because, in recent years, many works have shown that it is possible to segment the speech signal into syllable-like units using exclusively the acoustic properties of the signal [3][4].

Compared to the triphone, the syllable is a better choice because there are far fewer legal syllables than triphone patterns in a language and also because the syllable is more stable from an acoustic point of view [5]. Also, several papers [5] [6] have compared syllable and triphone based recognition systems and reported higher accuracies for the syllable based systems.

Although the syllable has for a long time been considered a suitable unit for speech recognition [7] there are few studies that deal strictly with syllable classification [8]. The approach used in [8] is based on a mapping between articulatory-acoustic features and syllables and it uses five articulatory-acoustic feature sets: manner, voicing, place, rounding and front-back. Each feature set was classified according to its own Artificial Neural Network (ANN) and mapped together in a syllable template. In order to obtain the syllables, a further classification based on these syllable templates was performed. The classification accuracy reported on the TIMIT corpus was 48.2%.

Several papers in the literature dealing with syllable-based speech recognition use syllable classification as an intermediate step. Sirigos et al. [9] constructed a hybrid system for syllable recognition based on ANN and Hidden Markov Models (HMM). Their main idea was to model the syllables using vowel spotting techniques. They obtained an accuracy of 75.09% for the TIMIT corpus and 59.30% for NTIMIT. In [10] Domont et al. interpreted spectrograms as images and deployed a hierarchical object recognition system. The authors report 0.9% WER with the integration into the system of a DTW procedure and 5.4% without the DTW, using a corpus made of 25 very frequent monosyllabic words extracted from the DARPA Resource Management database. Matsuoka et al. [11] perform syllable recognition using a method based on integrated neural networks. Grouping the syllables based on the manner of articulation of the consonants, they obtain 96.2% accuracy. Hu et al. [12] define syllable-like units by means of rules and represent them using trajectory models. The accuracy of their syllable units classification is 84.8%.

A similar approach to the one used in this work for feature extraction was proposed by Sarada et al. [13]. They make use of multiple frame size (MFS) and multiple frame rate (MFR) for feature extraction comparing ASR results with that of the single frame size (SFS), where the window size and frame rate are fixed. They observe a 4% improvement in the performance using the former method.

Ganapathiraju [14] proposes a hybrid system for speech recognition based on HMM and Support Vector Machines (SVMs). He shows that SVMs are well suited for speech recognition because they are based on the principle of "structural risk minimization" and they have a very high discrimination ability. In [15] Ganapthiraju et al. show that the use of SVMs provide an improvement in performance on a static pattern classification task based on phonemes. They also describe an application of SVMs to large vocabulary speech recognition, and obtain an improvement in error rate on a continuous alphadigit task (OGI Alphadigits) and a large vocabulary conversational speech task (Switchboard). Clarkson and Moreno [16] use SVM for phonetic classification and they show that SVM outperforms classification based on Gaussian mixtures on different data sets. On the TIMIT data set, they obtained an accuracy of 77.6% using SVM, representing an increase of 3.9% with respect to the Gaussian mixture.

Prosody has always proved itself to be important in disambiguating a number of controversial situations with

syllables. In [17], especially, phonemically identical sequences in monosyllabic and carrier words could be distinguished through prosodic boundaries detection.

Prosody has been used extensively in the recent years in speech related fields. Prosodic features extracted at the syllable level was used for different tasks, like: automatic stress detection [18], speaker recognition [19] and even language modelling [20]. Similar approaches to the one presented in this paper, but for speaker recognition [21] and speech recognition [22] showed that indeed the prosody adds complementary information to the cepstral coefficients.

This paper is organized as follows: section II introduces the methods used to carry out this work, namely SVMs, the syllable representation employed and the corpus used. Section III contains the details of the experiments conducted and the results obtained. We will conclude in section IV, while also introducing some future work.

# 2. Methods

## 2.1. Syllable representation

In this paper we aim at representing the syllable as a static analysis unit in analogy to what is done in the processes of feature extraction for face/image recognition algorithms. It means that all internal variations due to coarticulation or, more generally, to acoustic-temporal dynamics, are merged into a set of features that the SVM classifier will consider as a whole.

We explore in this study the following two approaches for syllable representation:

- using a fixed number of frames per syllable
- using a variable number of frames per syllable

In the first approach, each syllable was described using a fixed number of frames, *n*, regardless of the syllable length. In order for these *n* frames to cover the entire syllable while keeping the length of the analysis frame constant, the shift between two consecutive frames had to be varied from syllable to syllable.

Figure 1 illustrates this approach, where the numbers from *1* to *n* represent the frame number, while the number under each frame represents the distance between the beginning of two consecutive frames (being the sum of the frame length and the shift between the current frame and the next frame). It can be seen that for shorter utterances the frames overlap, while for longer utterances there are gaps between two consecutive frames.

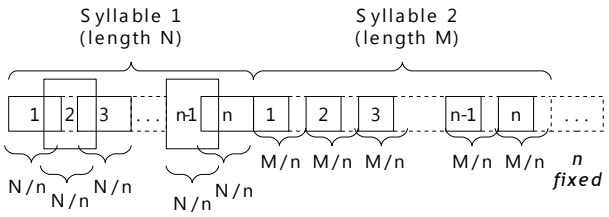


Figure 1. *Representation using a fixed number of frames per syllable*

The second approach uses a variable number of frames per syllables, the number being proportional to the length of the syllable. In this case, in order to cover the whole syllable length, the length of the analysis frame as well as the shift between consecutive frames were kept constant thus obtaining a variable number of frames per syllables, depending on the syllable length. This type of representation is exemplified in Figure 2. It can be seen that the distance between the beginning of two consecutive frames has the fixed value k, while the number of frames per syllable is a function of the length of the syllable.

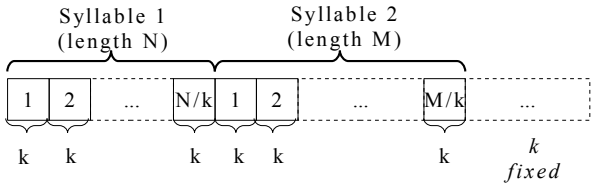


Figure 2. *Representation using a variable number of frames per syllable*

The second step consisted in choosing the set of features to be used. In the first part of the experiment, we used Mel Frequency Cepstral Coefficients (MFCC) extracted from 16 ms analysis frames using the facilities offered by the Auditory Toolbox [23]. The set of features was enriched, for the second part of the experiment, using basic prosodic features, like duration, energy and F0.

## 2.2. Corpus

The corpus we used for testing the procedure is a part of the Italian version of the SPEECON corpus [24]. It consists of 2106 files containing numbers between 0 to 999,999. The corpus was annotated at the syllable level, thus obtaining 10427 syllable occurrences, corresponding to the 42 different syllables present in the corpus. Because each file in the corpus contains silent periods at its beginning and end, we created for classification purposes an additional class, corresponding to the silence.

For our experiment we divided the corpus into three subsets: one for training, one for testing and a last one for development, such that no speaker appearing in one of the subsets would appear in the other two.

Because the class representing silences has significantly more members than any other class and because it is classified correctly in most of the cases, it was possible that it might bias the classification accuracy. In order to avoid this, we limited the number of members of the silence class for the three subsets to less than the maximum number of occurrences of any syllable in that subset.

# 3. Results

The experiments were divided in two parts, using: a) only MFCC and b) a combination of MFCC and prosodic features. Being a preliminary study, the procedure for syllable classification was applied to manually segmented data. To construct the SVM classifiers we used LIBSVM [25], a library for support vector classification, which is also able to perform multi-class classification.

## 3.1. Experiment 1

For the MFCC-based classification we took into account several experimental settings.

The initial idea was to divide the syllable in three equal parts (roughly equivalent to the onset, nucleus and coda) and to take from each part three frames equally distanced from each other. Afterwards we decided to expand the search for the best number of frames per syllable and so we tested the classification for 7, 9, 11, 13, 15 and 17 frames per syllable.

In order to be able to compare the results obtained with the for different number of frames per syllable, the best SVM parameters had to be found. The search for these parameters was done using a tool offered by the LIBSVM package, the *grid* function, which performs v-fold cross-validation. In our case, five folds were used on the training set. A preliminary test was conducted for finding the kernel best suited for this classification task. It confirmed that the best kernel is the RBF kernel and thus, this kernel was used in all the experiments.

The results of the classification tests for the six values for number of frames per syllable, using the RBF kernel, are presented in Table 1.

Table 1. *Classification accuracy using MFCC for fixed number frames per syllable*

| Number of frames / syllable | Accuracy [%] |
|---|---|
| 7 | 85.27 |
| 9 | 85.55 |
| 11 | 86.66 |
| 13 | 86.27 |
| 15 | 86.27 |
| 17 | 86.5 |

It can be seen from the results presented in the previous table that the syllable representation using 11 frames per syllables gives the best results in terms of the accuracy obtained and the compactness of the representation. A McNemar test was performed and it showed that the improvement in accuracy brought by the representation using 11 frames per syllables with respect to the previously best representation (9 frames per syllable) is statistically significant. Also, a further increase in the number of frames per syllable doesn't increase the accuracy.

The second type of representation of the syllable, using a variable number of frames per syllable, was tested for the following values of the shift between two consecutive frames: 128, 256 and 512 samples, corresponding to 8, 16 and 32 ms respectively. These cases would further correspond to an overlap of half a frame, no overlap and taking only one in two frames respectively. Table 2 presents the results obtained.

Table 2. *Classification accuracy using MFCC for variable number of frames per syllable*

| Frame shift [ms] | Accuracy [%] |
|---|---|
| 8 | 78.94 |
| 16 | 80.1 |
| 32 | 77.95 |

By comparing the maximum classification accuracies attained, it is clear that this syllable representation performs worse than the previous one. In light of the results presented in this section, we have chosen to use in the following experiment only the representation using a fixed number of frames per syllable, more exactly 11 frames per syllable.

### 3.2. Experiment 2

In the first part of this experiment we used a set of prosodic features for syllable classification while in its second part we tested whether the information carried by these prosodic features is complementary to the one carried by the MFCCs and to what degree the combination of the two types of features would improve the classification accuracy.

The set of prosodic features we used for representing the syllable had both global features, i.e. features having one value for the entire syllable, as well as more localized features, extracted for each analysis frame (11 frames/ syllable). The set of prosodic features consists of the following 23 features:

- duration (in seconds)
- 11 x F0 (in Hz)
- 11 x energy (in dB)

The classification results using the prosodic feature sets are presented in Table 3. It appears that they fail to discriminate between all the classes the syllables, due probably also to their low number of features.

Next, we tested if, by adding the prosodic information to the MFCC features would improve the classification accuracy. This was done by combining the class probabilities given by the classifiers of the two features according to:

$$P_{ic}=\alpha_i \cdot \log(P_{i1})+(1-\alpha_i)\cdot \log(P_{i2}), 0<\alpha_i \leqslant 1$$

where $P_{ic}$ represents the combined classification score of syllable $i$, $P_{i1}$ and $P_{i2}$ represent the probabilities of syllable $i$ given by the MFCC based and the prosodic based classifiers while $\alpha_i$ represents the weighting factor corresponding to the syllable $i$. The parameters $\alpha_i$ and were found using an exhaustive search on the development set, search aiming to maximize the accuracy of the combined system.

Table 3. *Classification results using MFCC and prosodic features*

| Representation | Accuracy [%] |
|---|---|
| MFCC | 86.66 |
| Prosodic | 43.59 |
| MFCC + Prosodic | 87.22 |

It can be seen that, by combining the information coming from the two classifiers, the classification accuracy has increased. Although it is a small increase, a McNemar statistical test showed that the increase is statistically significant.

When deciding to employ prosodic features for syllable classification we departed from the premise that the prosodic features would help in discriminating categories of syllable classes. In order to verify this idea, we grouped the syllables in the following eight categories, depending on their segmental structure: SIL, V, VC, CV, CVC, CCV, CCVC, CCCV. While the accuracy of the new syllable categories increased as expected, the absolute improvement obtained using the classifier based prosodic features, 23.97%, was far superior to the one obtained using the MFCC based classifier, 4.22%. This fact validates our opinion about the usefulness of prosodic features for discriminating between categories of syllables.

We also made an analysis of the results obtained with the prosodic based classifier and compared them with the results of the MFCC based system. We observed that the maximum improvement that the prosodic feature could bring to the MFCCs would be of 2.50%. The actual increase observed, 0.56%, is quite small compared to this maximum attainable value and it is due to the combination rule used. We think that a different decision making mechanism would bring a bigger improvement.

## 4. Conclusions and future work

At the present our system represents speech as a sequence of syllable-like units (in a time scale ranging from a minimum of ~70 to ~300 ms). This specific prosodic approach allows us to design a recogniser that can be based on time scales that are longer than that used in traditional systems. The use of an MFCC 11x13 matrix transforms temporal dynamics internal to the syllable into a snapshot of the unit considered as static, allowing us to substitute the classical HMM approach with a more efficient classification task.

The results obtained are encouraging, the accuracy from the use of the MFCC feature set alone is sufficiently reliable and consents the design of an appropriate decoding algorithm to calculate words and sentences probabilities by the superimposition of a language model. At the same time we would like to improve the classifier performances as we ideally aim at using syllable-like units deriving from an automatic syllabification process [4], presently in course of further improvements.

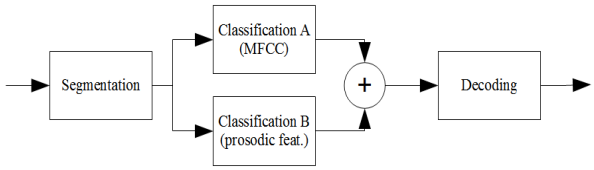


Figure 3. *Proposed system architecture*

In this view we think that the use of other prosodic features, inserted into a parallel (see Figure 3) architecture could provide us with a more robust classification performances. Even if the increment in the accuracy provided by adding the classification system A is presently below our expectations, we think that it is anyway encouraging and stimulates us to attempt new research in this direction, looking for other feature sets (we are currently experimenting Modulation Spectrogram [26]) and more sophisticated decision making techniques to be used to combine results deriving from the parallel classification branches.

## 5. Acknowledgements

Bogdan Ludusan's work was supported by the EU FP6 Marie Curie Research Training Network "Sound to Sense". The authors would like to thank Odette Scharenborg for her helpful suggestions and comments regarding the experiments conducted.

## 6. References

[1] D. Poeppel, "The analysis of speech in different temporal integration windows: Cerebral lateralization as asymmetric sampling in time", *Speech Communication*, vol. 41, pp. 245-255, Aug. 2003.

[2] M. Meyer, "The neural mechanisms of temporal acoustic information processing during speech perception", Invited talk, 5th AISV conference, 2009.

[3] T. Nagarajan, H. A. Murthy, and R. M. Hegde, "Segmentation of speech into syllable-like units", in Proceedings of Eurospeech 2003, pp. 2893-2896.

[4] M. Petrillo and F. Cutugno, "A syllable segmentation algorithm for English and Italian", in Proceedings of Eurospeech 2003, pp. 2913-2916.

[5] A. Ganapathiraju, J. Hamaker, M. Ordowski, G. Doddington, and J. Picone, "Syllable-based large vocabulary continuous speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol.9, pp. 358-366, May 2001.

[6] S. Nakagawa, K. Hanai, K. Yamamoto, and N. Minematsu, "Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition", in Proceedings of ASRU 1999, pp.393-396.

[7] S. Chang, "A syllable articulatory-feature and stress-accent model of speech recognition", Doctoral Thesis, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 2002.

[8] M. Wester, "Syllable classification using articulatory-acoustic features", in Proceedings of Eurospeech 2003, pp. 233-236.

[9] J. Sirigos, N. Fakotakis, and G. Kokkinakis, "A hybrid syllable recognition system based on vowel spotting", *Speech Communication*, vol. 38, pp. 427-440, Nov. 2002.

[10] X. Domont, M. Heckmann, H. Wersing, F. Joublin, and C. Goerick, "A Hierarchical Model for Syllable Recognition", in Proceedings of ESANN 2007, pp. 573-578.

[11] T. Matsuoka, H. Hamada, and R. Nakatsu, "Syllable recognition using integrated neural networks", in Proceedings of IEEE IJCNN 1989, vol. 1, pp. 251-258.

[12] Z. Hu, J. Schalkwyk, E. Barnard, and R. Cole, "Speech Recognition Using Syllable-Like Units", in: Proceedings of IEEE ICSLP 1996, vol. 2, pp. 1117-1120.

[13] G. L. Sarada, T. Nagarajan, and H. A. Murthy, "Multiple frame size and multiple frame rate feature extraction for speech recognition", in Proceedings of SPCOM 2004, pp. 592-595, 2004.

[14] A. Ganapathiraju, "Support Vector Machines for Speech Recognition" Doctoral Thesis, Department of Electrical Engineering, Mississippi State University, 2002.

[15] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition", *IEEE Transactions on Signal Processing*, vol. 52, pp. 2348-2355, Aug. 2004.

[16] P. Clarkson and P. J. Moreno, "On the use of support vector machines for phonetic classification", in Proceedings of IEEE ICASSP 1999, pp. 585–588.

[17] A.P. Salverda, D. Dahan, and J.M. McQueen, "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension," *Cognition*, vol. 90, pp. 51-89, 2003.

[18] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners", In Proceedings of ICASSP2005, pp. 733-736.

[19] L. Ferrer, E. Shriberg, S. Kajarekar, and K. Sonrnez, "Parametrization of Prosodic Feature Distributions for SVM Modelling in Speaker Recognition", In Proceedings of ICASSP2007, vol.4, pp. 233-236, 2007.

[20] S. Huang and S, Renals, "Using Prosodic Features in Language Models for Meetings", In Proceedings of MLMI 2007, pp. 192-203.

[21] L. Ferrer, E. Shriberg, S. Kajarekar, A. Stolcke, K. Sonmez, A. Venkataraman, and H. Bratt, "The Contribution of Cepstral and Stylistic Features to SRI's 2005 NIST Speaker Recognition Evaluation System", In Proceedings of ICASSP2006. vol.1, pp. 14-19.

[22] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Qin Jin, D. Klusacek, J.; Abramson, R. Mihaescu, J. Godfrey, D. Jones, and Bing Xiang, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition", In Proceedings of ICASSP03, vol.4, pp. 784-787.

[23] M. Slaney, *Auditory toolbox: Version 2*, Interval Research Corporation, 1998-010, 1998.

[24] R. Siemund, H. Höge, S. Kunzmann, and K. Marasek, "SPEECON - Speech Data for Consumer Devices", In Proceedings of International Conference on Language Resources and Evaluation (LREC), Athens, Greece, 2000, vol. 2, pp. 883-886.

[25] C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[26] S. Greenberg and B. Kingsbury, "The Modulation Spectrogram: In Pursuit of an Invariant Representation of Speech", in Proceedings of ICASSP97, vol. 3, pp. 1647-1650.