A SIMPLE PHONEME ENERGY MODEL FOR THE GREEK LANGUAGE AND ITS APPLICATION TO SPEECH RECOGNITION

Dimitris Tambakas, Iliana Tzima, Nikos Fakotakis, George Kokkinakis
Wire Communications Laboratory, University of Patras, 261 10 Patras, Greece
Tel: +30 61 991722, Fax: +30 61 991855, E-Mail:tambakas@wcl.ee.upatras.gr

ABSTRACT

This paper deals with the improvement of autosegmentation algorithms by establishing and implementing a simple energy model. This model consists of rules which describe the variation of the phoneme energy at the phoneme boundaries due to the phoneme context. The efficient estimation of phoneme boundaries results to the improvement of the accuracy of phoneme-based, large vocabulary speech recognition systems, as proven from experiments in the Greek language.

1. INTRODUCTION

Phoneme based speech recognition systems [1],[2] employ autosegmentation algorithms in order to extract the phoneme boundaries of candidate words from the speech signal. The expected phoneme duration, the distance between the signal and the phoneme prototype, the distance between the signal and the crossing point of successive phoneme prototypes, and the phoneme energy are the main parameters used in the autosegmentation algorithm. In the latter case, the phoneme energy level is usually considered, in order to detect energy transitions at the phoneme boundaries.

This paper presents an efficient model that estimates the boundaries of each phoneme string of a candidate word taking into account, a part of the phoneme energy, and the phoneme context, i.e the preceding and the following phoneme. This model has been established by analysing a large annotated speech data base of the Greek language. Statistical tests and ANOVA analysis were used to examine the contextual effects on the energy. The result of the analysis is a simple energy model consisting of rules which can estimate the phoneme boundaries with a statistical confidence of 95%. The validity of the presented model has been proven by testing it in the fine phonetic analysis of the Greek Isolated Word Speech Recognition (IWSR) system developed in the framework of the ESPRIT 2104 project POLYGLOT [3].

The paper structure is as follows: First, we describe the data bases used in this work. Then, we present the statistical methods applied in the data base analysis and the model we established. We also give a short overview of the IWSR-system and of the implementation of the energy model in it. Finally, the performance evaluation of the model is given.

2. DESCRIPTION OF THE SPEECH DATABASES

A 500-word speech database was used for the phoneme energy analysis. This database covers all the Greek phonemes and their most frequent contextual combinations. It contains words of various lengths with various syllabic structures in various locations. The 500 words were spoken in isolation by eight Greek native adult speakers (four male and four female). After that, the speech material was manually segmented and labelled. For the purposes of this work the energy at the phoneme boundaries was estimated.

The resulted phoneme database contains a total of 17,000 entries. Each entry consists of a) the current phoneme label, b) the word this phoneme belongs to, c) the phoneme's phonetic context (left and right), d) the speaker's name, c) the name of the speech file, and f) two fields with the energy level at the left and right phoneme boundaries.

3. SPEECH DATABASE ANALYSIS - THE ENERGY MODEL

The Greek phonemes were grouped into 7 categories according to the manner and the place of articulation (labial, dental, etc.). These categories represent the examined phoneme context.
For each phoneme its mean energy and the mean energy at its boundaries was measured for different phoneme contexts and the percentage change of the mean energy for the left and right context was calculated in each case.

Through ANOVA analysis and statistical tests on the whole data base the mean percentage change and the corresponding 95% confidence interval was calculated for each phoneme and context.

Table 1 gives as example these values for the right context of the phoneme /i/. Fig. 1 illustrates the mean energy (ME) and the confidence interval at the right boundary of /i/ for various right contexts.

Right context	Percentage of mean phoneme energy	Confidence interval
1. Labial	0.06	0.65
2. Dental	1.06	0.67
3. Fricative	0.70	0.44
4. Palatal	0.81	0.52
5. Liquid	1.29	0.99
6. Nasal	1.40	0.84
7. Vowel	1.45	1.06

Table 1: The right context rules for the energy of phoneme /i/.

The established values, like those of Table 1, represent the rules of the model for the estimation of the energy at the left or right boundary of a given phoneme. Only the contextual effects which create statistically significant changes in the mean phoneme energy are used as rules. In this way 150 rules were extracted as a whole.

The implementation of the rules provides an estimation for the position of the left and right boundary of a given phoneme according to its context. For example, the right boundary of the phoneme /i/, followed by a labial phoneme is in the interval (see table 1):

$$ME(/i/)x0.06 \pm 0.65$$

Some important findings from the analysis are the following:

1. The energy at the right boundary (rb) of each phoneme is affected significantly only by the right context (rc), while there is no important influence from the left context (lc). For example, let us examine the results of two sets of experiments for the phoneme /i/ by ANOVA analysis. The first set is extracted by changing the right context of /i/, the second by changing its left context. The results of the first set ($F_{rb/i/rc}=438.84$, $p<0.00001$) indicate that the energy is affected strongly by the right context but for the second set the effects from the left context are not of significant importance ($F_{rb/i/lc}=0.275$, $p=0.606$). Indeed, in Fig.1 we observe that the energy is significantly different for each context in the right boundary.

2. The same happens for the energy which is estimated at the left phoneme boundary. The effects of the left context are statistically significant, but the effects of the right context are not. For example, for the phoneme /i/ the left context affects significantly the energy in the left boundary ($F_{lb/i/lc}=47.55$, $p<0.00001$) but the right context does not affect the mean energy ($F_{lb/i/rc}=0.01$, $p=0.972$).

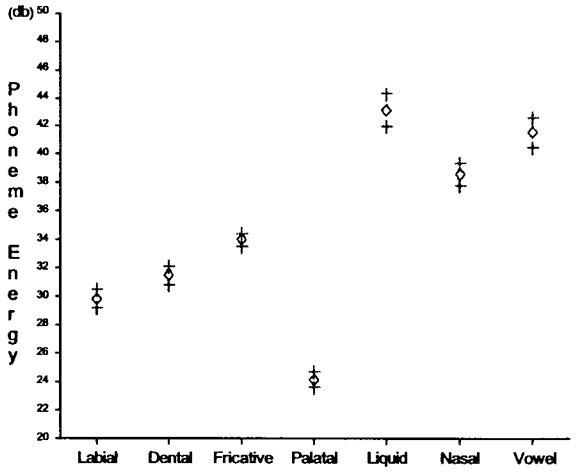

Figure 1: The mean energy and the confidence interval which has been calculated at the right boundary of phoneme /i/ for various right contexts.

3. The influence of the contextual environment is stronger on the vowel energy than on the consonant energy (e.g. for vowel /i/ and consonant /r/ we have estimated $F_{rb/i/rc}= 438.84 > F_{rb/r/rc}=14.028$, $p<0.00001$). The only exception is the consonant /s/ which is influenced by the context very strongly as it happens in the case of a vowel ($F_{rb/s/rc}= 471.98$, $p<0.00001$).

4. There are some consonants (/m/, /x/, /v/, /l/), the energy of which is not influenced by the contextual environment in contrast to vowels. For those consonants, ANOVA analysis proves that there is no statistically significant difference among the mean energies which have been estimated for each context (e.g. $F_{rb/m/rc}=0.275$, $p>0.606$).

According to the above mentioned results the energy model consists of two types of rules. The first type (rb/ph/rc) estimates the energy at the

right boundary (rb) of a phoneme /ph/ when the right context (rc) changes and we call it *right rule*. The second type (lb/ph/lc) estimates the energy at the left boundary (lb) of a phoneme /ph/ when the left context (lc) changes and we call it *left rule*. The implementation of the energy rules in the Greek IWSR-system is described in the next section.

4. IMPLEMENTATION OF THE ENERGY MODEL IN IWSR SYSTEMS

The aim of the energy model is to provide additional hypotheses for the phoneme boundaries during the Fine Phonetic Analysis (FPA) of phoneme based large vocabulary IWSR systems. Such a system for the Greek language and its FPA are briefly presented in the following subsections.

4.1. The Greek IWSR-system

The Greek IWSR-system consists of four modules: The **signal processing** module, that performs LPC analysis estimating 20 LPC and cepstrum coefficients for each 10 msec frame. The **phonetic recognition** module, where a rule-based phoneme recognition algorithm constructs a tentative phoneme string, without using lexical information. The **preselection** module, where a smaller set of words (about 40 words) is selected from the entire vocabulary (8000 words). This is achieved by finding out the similarity between the previously extracted string and each word of the dictionary. The similarity is estimated by Dynamic Programming string matching. Finally, the **fine phonetic analysis** module, that gives a set of 1 to 5 candidate words. These words are the most similar to the signal out of the initial set of 40 words and are employed by the subsequent language model.

4.2. Fine phonetic analysis and the energy model implementation

The FPA enables the system to find out the best alignment (definition of phoneme boundaries) between each word of the dictionary and the parametric description of speech.

To do that, for each phoneme, it examines a set of frames as possible right boundaries, which we call hypotheses. These frames are selected using two criteria:

1. The crossing point of successive phoneme prototypes. For example, suppose that there are two successive phonemes ph_1 and ph_2. Two curves are created from the distances between each phoneme prototype [ph_1], [ph_2] and the signal frames. Then, the crossing point of the two curves is a hypothesis for the limit between the two phonemes ph_1 and ph_2.

2. The energy transitions between the successive phonemes. For example, if the mean energy of phoneme ph_1 is higher than the energy of the phoneme ph_2, then an increment of the transition of energy is observed. This transition defines an hypothesis for the boundary between the two phonemes.

If there are not hypotheses from the previous criteria then the mean phoneme duration is used to define a possible phoneme boundary (default hypothesis). The criteria can not produce hypotheses for two reasons. First, when the phoneme prototype is not robust and the distances between this phoneme prototype and the signal frames are not reliable. Second, when the mean phoneme energy of the consecutive phonemes is approximately in the same level.

Each hypothesised limit is phonetically evaluated using local scores. The scores are generated for each limit, comparing the phoneme duration and the distance from its prototype with the corresponding standard values. These values are taken from histograms derived from a large phonetically labelled speech database which is representative of many speakers. The hypotheses with the highest score, which equivalently mean a high value of the maximum likelihood function, are accepted, while the others with very low score are discarded. The scores for each accepted phoneme limit and for all the phonemes of the word are added together on an logarithmic scale.

A left-to-right beam, created from the accepted hypotheses of each phoneme, is used to find the alignment score. The hypotheses for the phoneme boundaries, from which the best alignment score is calculated, make up the final phoneme boundaries.

The FPA algorithm leads to less accurate results when there are not any hypotheses from the criteria for the phoneme limits and only the default hypothesis is used. This paper proposes the

employment of the described simple energy model in order to acquire additional hypotheses for the phoneme boundaries.
The new hypotheses concerning the limit between two phonemes, are created using the left and right energy rules as follows:

1. The right energy rule is used in order to detect the right boundary for the first phoneme. The algorithm searches for the frames, the energy level of which is within the confidence interval of this rule. The searched area is defined from the minimum and maximum phoneme duration.

2. Using the same approach, the left energy rule is used in order to find additional hypotheses for the boundary between the phonemes. The left energy rule creates new hypotheses for the left limit of the second phoneme. The implementation of the left energy rule is useful especially in the case where the right energy rule is not robust (i.e. when the first phoneme is one of the /m/, /x/, /v/, /l/).

If there are many frames complying with the right and left rule then the algorithm provides only the best. The frames which are considered to be the best, are those that are close to the mean energy of the rule. The frame which results from the best frame of the right rule and from the best frame of the left rule and is situated in the middle is also considered.

5. PERFORMANCE EVALUATION

To evaluate the performance of the described energy model we compared the recognition results employing the described IWSR-system both with the energy model and without it. The recognizer used a 8,000 words vocabulary. Table 2 shows the results for five speakers and up to five candidate words. The order of the words gives the number of candidate words in which the correct word is included and the corresponding success rate. The recognition performance has been improved significantly, when the energy model was used. Table 2 shows this improvement. For all speakers it is greater than 16% for one candidate word and greater than 2% for five candidate words, while improvement as large as 25% has been observed in separate cases.

Cand.	Without Energy Model				
Word	SP1	SP2	SP3	SP4	SP5
1	43.65	40.50	52.23	53.10	43.94
2	50.76	48.25	68.15	59.24	56.48
3	63.02	61.12	79.63	69.72	76.93
4	83.56	82.75	83.71	77.44	81.71
5	94.01	90.05	91.63	88.54	90.99
Cand.	With Energy Model				
Word	SP1	SP2	SP3	SP4	SP5
1	60.11	59.55	69.63	70.54	60.06
2	76.63	72.23	81.11	78.29	68.78
3	88.15	84.15	87.40	86.04	85.57
4	90.56	89.92	92.22	89.92	87.49
5	96.10	92.58	94.81	91.86	93.50

Table 2: Success rate (%) of the IWSR-system for five speakers (SP1-SP5) and up to five candidate words

6. CONCLUSION

In this paper we described the establishment and the implementation of a simple context dependent energy model. This model is composed of 150 rules and provides additional information on the phoneme boundaries. Using this model for a phoneme based, large vocabulary speech recognition system, the recognition accuracy is improved.
In order to evaluate this improvement, we tested a Greek IWSR-system both with an energy model and without it. In the first case the results were at least 2% and up 25% better than in the second case.

6. REFERENCES

[1] R. Billi, G. Arman, D. Cericola, G. Massia, M.J. Mollo, F. Tafini, G. Varese, V. Vittorelli (1989), "A PC Based Very Large Vocabulary Isolated Word Speech Recognition System", Eurospeech 89 (Paris), Vol 2, pp 157-160.

[2] P. Buttafava, R. Billi, W. Diagiampietro Massia V. Vittorelli (1989), "Architecture and Implementation of the Olivetti PC-based Very Large Vocabulary Isolated Word Speech Recognition System" Eurospeech 89 (Paris), Vol 1, pp 90-93.

[3] D. Tambakas, N. Fakotakis, G. Kokkinakis (1995), "Robust Phoneme Prototype extraction for Speech Recognition" Eurospeech 95 (Madrid), Vol 2, pp 927-930.