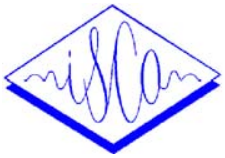

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

SELECTION OF THE MOST EFFECTIVE SET OF SUBWORD UNITS FOR AN HMM-BASED SPEECH RECOGNITION SYSTEM

A. Tsopanoglou[1], *N Fakotakis*[2]
[1]KNOWLEDGE S.A., Human Machine Communication Dept.,
N.E.O Patron Athinon 37, 264 41 Patras, Greece
Tel: +30 61 452820, Fax: +30 61 453819, e-mail:KNOWLEDGE@Patra.hol.gr

[2]Wire Communications Laboratory (WCL), Electrical & Computer Engineering Dept.,
University of Patras, 261 10 Patras, Greece
Tel: +30 61 991722, Fax: +30 61 991855, e-mail:fakotaki@WCL.ee.upatras.gr

10.21437/Eurospeech.1997-31

ABSTRACT

In this work we describe several approaches to determine an effective set of subword units for modeling the spoken Greek language. We tried to form a concrete set of basic units which must have the capability of giving a unique phonetic transcription for every input utterance. The results of an extensive set of experiments showed that the use of longer units than phonemes can lead to a significant improvement in a system's performance. Three sets of subword units were finally formed regarding the way we combined the 42 phonemes of the Greek Language. The three approaches showed better results than the baseline phoneme-based system and the most effective one proved to be the second approach in which we used two-phoneme combinations of the types non-vowel/vowel and non-vowel/non-vowel. The phoneme recognition rate of the system increased almost by 9% (reaching a level of 78.65%) for the best situation compared to the baseline system.

1. INTRODUCTION

The modeling technique (HMM, TDNN, etc), the recognition algorithms and the choice of the basic recognition units are the most critical factors of a speech recognition system. The choice of these parameters influences the performance and the functionality of the final system. Concerning the recent developments in the speech recognition field, we realize that the problems of choosing modeling techniques and decoding algorithms have been totally or partial solved. The Hidden Markov Modeling and the efficient decoding algorithms such as Level Building and Frame Synchronous Viterbi Beam Search [2],[3],[4] seem to give a complete solution to the recognition problem. However, an essential factor which remains undetermined and is open to debate is the choice of the basic recognition units which varies according to the specific application, to the used language or to the special characteristics of the hardware platform on which a system is installed. Several approaches have been used during the last two decades, beginning from the use of words and ending up to the use of subword units[4],[5],[6]. The word based systems have been proved to give excellent results when they are used in a specific application with small and fixed vocabulary. But their inherent difficulty in being easily adopted to other applications forced the researchers to use different more flexible approaches such as phonemes and subword units.

The phonemes seem to give a good solution to the problem of building lexicon independent speech recognition systems which can work and can easily be adopted in many different applications. The performance, however, of the phoneme based systems proved to be poorer than the word-based ones especially because of the significant effect of the coarticulation on the recognition performance.

The intermediate solution of using units shorter than the words and longer than the phonemes (syllables, triphones, diphones, etc), seems to be the best solution to the aroused problem. A set of carefully extracted subword units can transcribe every vocabulary word and can improve the recognition performance of the system by giving the ability of modeling the between phonemes coarticulation effects. Many approaches of using subword units have reported to the literature. The generalized triphones, the syllables or syllable like segments, the diphones etc are some of the most popular units within the speech recognition community. The rules of choosing the most appropriate set of units are not yet concrete and many different approaches have been presented by the international literature.

In an attempt to increase the performance of our baseline phoneme-based continuous speech recognition system we used different sets of recognition units. By combining the 42 phonemes of the Greek Language in three different ways we obtained three sets of subwords units (SET1, SET2 and SET3) and we tested the influence of each set on the baseline system. The choice of the most effective set was based on the final performance of the system, on the computational and time requirements and on the memory requirements of the system.

The contents of each set of the subword units will be described in the section 2.1, the parameters of the recognizer are described in section 2.2, and within section 2.3 a brief description of the speech database used for the training and the testing of the system is given. The

recognition results are showed and explained within section 3 and the conclusions of our experimental work along with some indications concerning future work are given in section 4.

2. SYSTEM DESCRIPTION

The presented system focuses on the choice of the recognition units and so the most of the following description refers to the formation of the three sets of the basic units and on their influence on the final performance of the speech recognition system.

2.1. Subword Units

As we mentioned above we used three sets of syllables, where the word "syllable" corresponds to every valid combination of a group of non-vowel phonemes and a vowel.

The SET1 of the subword units contains 879 items which are the valid syllables of the Greek language and some additional items for the silence and the Greek ending phonemes (all vowels and /s/,/n/). This set consists of:

- the 101 two-phoneme syllables of the type non-vowel/vowel,
- the 529 three-phoneme syllables (two non-vowel phonemes and one vowel),
- the 226 four-phoneme syllables (three non-vowel phonemes and one vowel),
- the 7 five-phoneme syllables (four non-vowel phonemes and one vowel),
- the 10 two-phoneme combinations of the type vowel/non-vowel (vowel+ending phoneme),
- the 5 vowels of the Greek Language,
- 1 model for the silence.

The SET2 contains 294 phoneme combinations that can be used to give a unique phonetic transcription for every vocabulary word. The contexts of the SET2 are:

- the 101 two phoneme syllables of the type non-vowel/vowel,
- the 177 two phoneme combinations of the type non-vowel/non-vowel,
- the 10 two-phoneme combinations of the type vowel/non-vowel (vowel+ending phoneme),
- the 5 vowels,
- 1 model for the silence.

The SET3 contains only 170 models. Special care was given for the vowels phonemes and the combinations formed choosing non-vowel phonemes in the neighborhood of the vowels. The contents of the SET3 are:

- 101 two phoneme syllables of the type non-vowel/vowel,
- 40 two phoneme clusters of the type vowel/non vowel,
- 10 two-phoneme combinations of the type vowel/non-vowel (vowel+ending phoneme),
- 18 phonemes (5 vowels and 13 non-vowels) which can not be always included in bigger combinations),
- 1 model for the silence

We have to mention here that all the phonemes combinations were extracted from a large text corpora (over 20,000,000 words) which had been collected by the Wire Communications Laboratory. Table 1 summarizes the contexts of the three sets of the subword recognition units.

Combinations	SET1	SET2	SET3
1NV/V	**101**	**101**	**101**
2NV/V	**529**	-	-
3NV/V	**226**	-	-
4NV/V	**7**	-	-
NV/NV	-	**177**	-
V/NV	**10**	**10**	**50**
Phonemes	-	-	**13**
Vowels	**5**	**5**	**5**
Silence	**1**	**1**	**1**
Total	879	294	170

Table 1*. Contexts of the three sets of subword units which used by the speech recognition system (NV = Non-Vowel Phonemes, V = Vowel phonemes).*

2.2. Recognition System Parameters

The Hidden Markov Models have been proved to be the most efficient way of modeling the speech signal. The structure of the model, the number of states per model and the representation of the acoustic parameters vary according to the length and the nature of the recognition units. Our system involves a left-to-right Continuous Densities Hidden Markov Model for each phonemic combination. Each model has three states (N=3) and the description of the speech observations for each state is given by three Gaussian mixtures (M=3). Special attention was given to the incorporation of the unit duration and energy probabilities in the model.

The segmental k-means algorithm has been used for the estimation of the models' parameters such as the transition propabilities between the states, and the mean and covariance vectors of each Gaussian mixtures and each state. The five step training procedure follows [1],[7]:
Initialization: Every sentence is linearly segmented into a number of sections that amounts to the number of phonetic units representing the spoken sentence. In addition, each segment is linearly apportioned among the states of the corresponding model.

Clustering: The k-means algorithm is applied to the speech segments corresponding to each HMM state to provide M clusters (mixtures).

Model Estimation: The mean vectors and the diagonal covariance matrices of each cluster are estimated assuming

the clusters are described by Gaussian probabilities densities. Along with the mean and covariance vectors (μ_{mj} and U_{mj}; m=1,M, j=1,N), we estimate the transition state probabilities (a_{ij}; i=1,N, j=1,N) and the weights c_{mj}, for each model as the number of frames belonging within the m^{th} cluster.

Segmentation: The newly estimated models are used to re-segment each training sequence using a Viterbi decoding algorithm and the phonetic transcription of the utterance.

Algorithm Iteration: The steps Clustering, Model Estimation and Segmentation are iterated until the final converge of the algorithm. In our implementation the training procedure is completed after 5 iterations.

The Level Building algorithm was used for the recognition stage and for the segmentation phase during the training procedure[4],[7].

The speech signal was selected via a omni-directional microphone and was digitized at a sampling rate of 16kHz and stored on a hard disk of a PC. The LPC-based cepstral representation of each sentence was also stored in advanced. The front-end processor produces a vector of 19 cepstral coefficients from the 19 linear prediction coefficients, using a hamming window of 20msec and a frame step of 10 msec. We also used the normalized log energy of each temporal frame by involving a dynamic normalization scheme. The differential cepstral coefficients and the differential log energy are also added in the parameter vector in order to include a precise description of the dynamic features of the speech signal.

2.3. Speech Databases

A large text corpora was used in order to achieve a real coverage of the appearance of each phonemic combination in the Greek language. This corpora contains more than 20,000,000 words and has been produced by the Wire Communications Laboratory. The lexicon of the corpora is bigger than 200,000 words and covers many different social and scientific domains. The set of the extracted words consists of 1,500 different words which contain all the valid phonemic combinations. The set of the 1,500 words was separated into 10 sentences with 150 words each.

The acquisition of these sentences was carried out in an office environment, using two different PCs with different equipment (different audio cards etc) by ten male speakers. Each speaker uttered once the speech material. The final speech database, consisting of 15,000 was separated into two parts. The first part contains 7,500 words (5 sentences for each speaker) and was used for the training of the system and the second part which contains 7,500 words, was used to test the performance of the system.

In addition, an extra speech database, which contains a set of different words and was acquired three months ago, was added to the test speech database. This database consists of 200 words containing digits, letters, names of months, dates and frequent command words and each word has been uttered twice by ten speakers in isolation mode giving a total amount of 4,000 words. So, the final testing speech database consists of 11,500 words.

3. EVALUATION RESULTS

The overall performance of the system is measured by counting all the possible recognition errors (insertions, deletions and substitutions). The ***phoneme recognition rate*** PRR defined as [1]:

$$PRR = 100\left(1 - \frac{ins + dels + subs}{correct}\right) \qquad 3.1$$

where *subs, ins, dels* and *correct* are the number of substituted, inserted, deleted and correct recognised phonemes, respectively, evaluated by aligning the phonetic output of the recognizer with the correct phonetic sequence through a dynamic programming procedure.

Several tests were carried out in order to verify the system's recognition rate, PRR, and to define the influence of each phonemic combination on the system's performance. The three different approaches proved to give better results than the baseline phoneme based speech recognition system.

During the first test we used the SET1 and we observed an improvement in the performance of the overall system of more than 4%. On the other hand, a significant increase in the insertions rate was also pointed out. The main reason for the previous mentioned increase was the use of large phonetic cluster where an insertion of only one phonetic unit causes an insertion of up to 5 phonemes. Another drawback of using all the possible Greek syllables is the enormous increase in the system's memory (5.1 Mbytes) in order to store the model's parameters and increase computational requirements. We noticed that more than 45 seconds were needed in order to decode a speech signal of one second length on a Pentium PC (CPU at 120 MHz).

The second set of units (SET2), contains fewer than the half units of the SET1 and this difference forced the insertion rate and memory requirements to be significantly reduced. A significant decrease in the insertion rate was observed which led to an increase in the overall recognition rate of the system. The system's recognition rate for the SET2 reached the level of 78.65% and the required response time reduced by a factor of 4 (Table 3).

The last experiment concerned the use of the SET3. It was proved that without loosing much of the recognition rate we managed to reduce the memory and computational time requirements. This development led to a system with acceptable performance and low memory and

computational time requirements. Even if the system's requirements are significantly reduced compared to all the other approaches the final performance does not meet our standards and we are not in favour of using the contents of the SET3 as basic recognition units.

Table 2 gives the phoneme recognition accuracy versus the iteration number of segmental k-means algorithm. An improvement of more than 10% is observed between successive iterations while the iteration number is smaller than 3 but a slight improvement is achieved for every loop after the third iteration.

Phoneme Recognition Rate (%)					
Iterations	1	2	3	4	5
Baseline System	**21.19**	**35.73**	**54.76**	**59.12**	**68.67**
SET1	**25.12**	**43.18**	**66.76**	**69.32**	**73.89**
SET2	**30.53**	**45.61**	**70.06**	**76.80**	**78.65**
SET3	**29.56**	**44.98**	**67.83**	**72.54**	**74.92**

Table 2: *Recognition Rate for each set of recognition units according to the iteration number of the Segmental k-means algorithm.*

Table 3 shows the computational time and the memory requirements concerning each approach. The memory space refers to the required space for the storage of model's parameters (transition probabilities, mean and variance vectors for each mixture) and not to the storage requirements of the recognition algorithm. The first column gives an estimation of the required time from the recognition algorithm to decode a second of speech input.

	Time in sec for 1 sec input Speech	Memory Requirements (Mbytes)
SET1	**45.4**	**5.12**
SET2	**12.7**	**1.71**
SET3	**5.1**	**0.99**

Table 3: *Memory and computational time requirements of the speech recognition system for each experiment.*

	Insertions	Deletions	Sub/tions	Correct
Baseline System	**210**	**98**	**155**	**1478**
SET1	**272**	**51**	**87**	**1535**
SET2	**223**	**54**	**73**	**1566**
SET3	**258**	**55**	**85**	**1550**

Table 4: *Number of the correct recognized phonemes and of the recognition errors for four experiments using different recognition units.*

Finally, Table 4 gives some results concerning the phoneme recognition rate of the system in proportion to the used set of subword units. These results have been extracted using a small amount of test speech data which contained only 1693 phonemes and are given for forming a correct opinion about the distribution of the deletion, insertion and substitution errors.

4. CONCLUSIONS

Even if the presented results are very preliminary, a valuable conclusion haa been made concerning the characteristics of the final speech recognition system and the set of the subword units. The use of the SET2, consisting of the two phoneme combinations (Table 1) gives a good compromise between the performance of the system and the requirements (memory and computational) of the system.

Taking into consideration the extracted results we expect a significant improvement on the results with only few alterations. Some alterations may concern the diminish of the phonemic clusters' number in order to reduce the time required during the recognition stage. On the other hand, very valuable results will be given using the huge speech database which is currently being gathered within the EEC project SPEECHDAT II (LE2-4001) in which the KNOWLEDGE S.A. and the Wire Communications Laboratory are participating.

5. REFERENCES

[1] R. Pieraccini, *"Speaker Independent Recognition of Italian Telephone Speech with Mixture Density Hidden Markov Models"*, Speech Communication, Vol. 10, pp. 105-115, 1991.

[2] H. Ney, *"The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition"*, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP 32, NO. 2, April 1984.

[3] C.H Lee, L.R. Rabiner, *"A Frame-Synchronous Network Search Algorithm for Connected Word Recognition"*, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 37, NO. 11, November 1989.

[4] L.R. Rabiner, S.E.Levinson, *"A Speaker-Independent, Syntax-Directed, Connected Word Recognition System Based on Hidden Markov Models and Level Building"*, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-33, NO.3, June 1985.

[5] R. Schwartz, Y. Chow, O. Kimball, S. Krasner, J., Makhoul, *"Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech"*, IEEE International Conference on Acoustics, Speech and signal Processing, April 1985.

[6] K.F. Lee, *"Automatic Speech Recognition: The Development of the SPHINX System"*, Kluwer Academic Publishers, Boston 1989.

[7] L.R. Rabiner, J.G. Wilpon and B.H. Juang, *"A Segmental k-Means Training Procedure for Connected Word Recognition Based on Whole Word Reference Patterns"*, AT&T Technical Journal, Vol. 65, No. 3, pp. 21-31, 1986.