WORD VERIFICATION USING CONFIDENCE MEASURES IN SPEECH RECOGNITION

M.C. Benítez, A. Rubio, P. García and J. Diaz Verdejo

E-mail `carmen@hal.ugr.es`
Dpto. de Electrónica y Tecnología de Computadores
Universidad de Granada, 18071 GRANADA (Spain)

ABSTRACT

In this work we propose a novel way of discriminating the words that are recognized by a speech recognition system as correctly or incorrectly detected words. The procedure consists of the extraction of a set of characteristics for each word. Utilizing these characteristics, we have built two classifiers: the first one is a vector quantizer, while the second one, though also a vector quantizer, was trained using adaptative technique learning. The results obtained show an improvement in the performance of the recognizer achieved by reducing the number of insertions with no significant reduction in the correctly detected words.

1. INTRODUCTION

There are some applications of automatic speech recognition (ASR) in which it is not necessary for the system to recognize exactly all the words that appear in one sentence in order to provide an adequate response. This can be generated by extracting some of the information that it is carried by the input acoustic sequence and is usually called wordspotting in the literature.

Another interesting problem is the building of task independent wordspotting systems that are easily adaptable to extract different sets of keywords. Within a wordspotting system, the voice to be recognized is classified into words inside and outside the vocabulary. To model the keywords, subword units are used; the word model is obtained by the concatentation of the subword models. The models for the words outside the vocabulary are treated using different approaches from a small number of filler models to all possible words that may appear in the context of the keywords. In the bibliography [1] it is shown that an increase in the number of words considered as filler models improves the recognition rate and decreases the number of false alarms. However, computing time also increases with the number of filler models considered.

On the other hand, as present recognition systems are far from being perfect, it is necessary to define an estimation of the confidence in the hypothetical words being correct or incorrect. Many investigators have focussed their efforts in this direction. In this work, we design and describe a classifier to discriminate the hypothetical words into correct or incorrect. The procedure consists of classify a string of putative words into the correct words class C or the incorrect words class I. For this purpose, and due to the reasonable percentage of phonematic recognition provided, we shall use the information contained in the string of phones generated by the recognizer when a phoneme bigramam is utilized. This information will be combined with the ouput information from the recognition system for every detected keyword when a loop grammar is considered, and the keyword models compete with the filler models. This enables us to build a vector of characteristics for each putative word.

In Section 2, we describe the baseline system and the database used for training and testing. Section 3 presents the selection of characteristics used in the classification. In Section 4, the design of the classifier is shown and in Section 5 we introduce the confidence measures used in this work. In Section 6, the experimental results are discussed and, finally, in Section 7 we shall comment on the defects of the presented method and propose possible solutions.

2. THE BASELINE SYSTEM

The reference system has been developed by the members of GIPSyC[2]. The system uses semicontinuous hidden Markov models (SCHMM). The voice signal is sampled at 16 kHz. The parametrization process provides 14 cepstrum parameters in MEL scale and its first and second derivatives and also the energy and its first and second derivatives.

As we seek to design a task–independent wordspotter, we have used as a database for model training a set of 1400 sentences. These sentences were pronounced by 74 different speakers and are phonetically balanced. The test database is composed of 600 sentences emitted by 12 different speakers which are not included in the training database. These sentences are related to Spanish geography and are part of the Albayzin geography database [3]

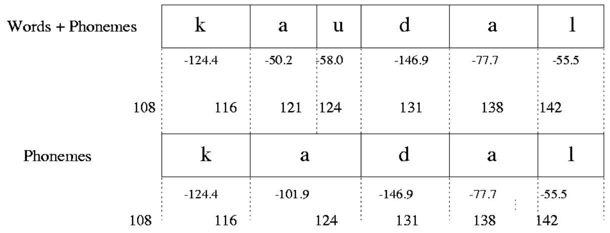

Figure 1: Alignment between the detected word and the sequence of recognized phonemes. The number at the beginning of the frame is the starting time and the one in the middle is the probability of each phoneme

The basic acoustic unit of recognition is provided by the independent context phones; for Spanish there are 23 different phones, although this number can be increased if alophonetic variations are considered. Within this work, these variations have not been taken into account; additionally we have constructed a background model for the silence. The set composed of the 24 models has been used not only for modelling the keywords but also for the words out of the vocabulary. The keyword models are obtained by the concatenation of the models indicated by their phonetic transcription. A loop grammar with the keyword models and the 24 models as filler models has been used. In order to obtain a higher recognition rate, the gramatical transtions for the keywords and the filler models have been weighted using factors of 1.8 and 1.6. This implies that we have not considered any information about the context in which keywords appear.

In the recognition process the search space is explored by using the Viterbi algorithm to obtain the best sequence of keywords and phonemes. In a parallel form, a phonematic recognition process is performed using a phone bigramam obtained from the training sentences.

3. EXTRACTION OF CHARACTERISTICS

In order to extract the information to be used in the classification process, we build a vector of characteristics for each hypothetical keyword. For each detected word, a time–alignment is performed with the corresponding sequence of phonemes obtained in the phonematic recognition process. An example of this alignment is shown in Figure 1. From this process we obtain a vector with 7 components which are:

• n_f. Number of phonemes of the hypothetical word.

• Logarithm of the probability normalized to the number of frames and phonemes

$$\frac{1}{n_f}\sum_{i=1}^{n_f}\left(\frac{1}{n_{f_i}}\sum_{j=1}^{n_{f_i}}\log P(o_j|\lambda_j^W)\right) \quad : \tag{1}$$

where n_{f_i} is the number of frames in the i-th phoneme, o_j is the acoustic vector of the j–th frame of the word and λ_j^W represents the state of the model provided by the Viterbi algorithm for the j–th segment.

• The third component is

$$\frac{1}{n_f}\sum_{i=1}^{n_f}\left(\frac{1}{n_{f_i}}\sum_{j=1}^{n_{f_i}}\left|\log P(o_j|\lambda_j^P)-\log P(o_j|\lambda_j^W)\right|\right) \tag{2}$$

where λ_j^P is the model state obtained by the Viterbi algorithm for the j–th segment for phonematic recognition.

• The fourth component is

$$\frac{1}{n_f}\sum_{i=1}^{n_f}\frac{1}{n_{f_i}}\left(\sum_{j=1}^{n_{f_i}}\left|\log P(o_j|\lambda_j^W)-\log P(o_j|\lambda_{H_i}^P)\right|\right) \tag{3}$$

where $P(o_j|\lambda_{H_i}^P)$ is the probability that the i–th phoneme finishes at the j–th frame.

• The fifth component is

$$\frac{1}{n_f}\sum_{i=1}^{n_f}\frac{1}{n_{f_i}}\left(\sum_{j=1}^{n_{f_i}}\left|\log P(o_j|\lambda_j^W)-\log P(o_j|\lambda_{H_{max}}^P)\right|\right) \tag{4}$$

where $P(o_j|\lambda_{H_{max}}^P)$ is the maximum of the probabilities that any of the phonetic models finishes in the j–th frame.

• The sixth component is

$$\frac{1}{n_f}\sum_{i=1}^{n_f}\frac{1}{n_{f_i}}\sum_{j=1}^{n_{f_i}}\mathrm{dist}_{Kull}\left(\lambda_{H_j}^P,\lambda_{H_j}^W\right) \tag{5}$$

where

$$\mathrm{dist}_{Kull}\left(\lambda_{H_j}^P,\lambda_{H_j}^W\right)=\sum_{i=1}^{N_e}\sum_{j=1}^{N_s}\left(P(o_j|\lambda_{H_i}^P)-P(o_j|\lambda_{H_i}^W)\right)$$
$$\log\left(\frac{P(o_j|\lambda_{H_i}^P)}{P(o_j|\lambda_{H_i}^W)}\right) \tag{6}$$

where N_e is the number of states and N_s the number of symbols

• The seventh component is

$$\frac{1}{n_f}\sum_{i=1}^{n_f}\frac{1}{n_{f_i}}\sum_{j=1}^{n_{f_i}}\delta_{\lambda_{H_j}^P,\lambda_{H_j}^W} \tag{7}$$

where $\delta_{k,l}$ is the Kroenecker delta and $\lambda_{H_j}^X$ stands for the phoneme obtained by the Viterbi algorithm for the j–th frame.

As can be seen, we have not included any grammatical information within the characteristics, which have been built only using the acoustic probabilities. The physical meaning of the two first parameters is obvious. The third parameter quantifies the acoustic similarity within the time the keyword is generated between this and the phonematic Viterbi sequence. The fourth parameter quantifies the differences between the phonemes probabilities in each keyword and the probability that these phonemes finish at the same time when phonematic recognition is used. If the most probable phoneme is used instead the same phone in the keyword, the fifth parameter is obtained. The last two parameters compare the transcription of every keyword to the phonemes obtained with the Viterbi sequence.

4. CONSTRUCTION OF THE CLASSIFIER

The problem proposed involves the discrimination between two classes, the class of the correct words (C) and the class of the incorrect ones (I). The classifier was implemented by using the self organizing map as pattern classification. The original map was obtained from vector quantization [4]. For the implementation of the classifier, a set of vectors of known class is needed in order to obtain the original dictionary. We used the training database to obtain these vectors. To generate the class C vectors, we generated a finite state automaton with the text of the sentences in the training database and this was also used in the recognition process. Making a comparison of the output string of words with the transcription of the spoken text, we tagged the vector of the characteristics of each word as C or I. In this way, a high number of members of C class were generated but a low number of members of I class. In order to increase the number of members in I, we selected a group of words from all the ones appearing in the training sentences and generated a loop grammar to be used in the recognition process. By repeating the tagger process, the members in the I class were increased.

With the tagged vectors, an initial dictionary is obtained, the assignation of tags to the centres of the dictionary is performed by majority vote among all the input vectors assigned to every centre. Every vector of unknown class is identified with the nearest vector code and is classified in the same class as the corresponding vector code. We used the euclidean distance to establish the closest vector. In order to improve the classification process, we applied a technique of adaptative learning using a LVQ algorithm to the initial dictionary [4].

5. WORD CONFIDENCE METRICS

In this section, we present the different quantities that we shall use to show the improvements caused by the classification procedure. Different confidence measures have been presented in the literature [5, 6, 7, 8].

One of the most frequently quantity used in the literature is the cross entropy (CREP) defined in [7] as:

$$CREP = \frac{1}{N} \sum_{w} \left[\delta_w \log(c_w) + (1 - \delta_w) \log(1 - c_w) \right], \tag{8}$$

where c_w is the probability that the hypothetic word is right for the given set of observations relatives to the word w and δ_w is 1 (0) for a correct (incorrect) word. If CREP obtained after the classification process is larger than the previous one we would infer that our confidence predictions are better . We used CREP in an averaged way in the same way as in [8].

Other quantities that indicate the quality of the classification procedure are the relative reduction in the number of insertions and the relative increase in the rejected keywords. We call them relative as they are a comparison of the results obtained before and after the classification procedure.

	Class C		Class I	
Component	Mean	Variance	Mean	Variance
2	-12.13	4.38	-13.05	5.00
3	0.67	0.95	1.14	1.15
4	11.73	6.91	12.78	8.07
5	34.17	36.66	35.36	41.82
6	5.59	5.04	8.70	5.45
7	0.24	0.20	0.40	0.23

Table 1: Mean and variance of the components of the characteristic vector for the classes C and I. The first component is the number of phonemes.

6. RESULTS

In this section we present the experimental results obtained with the techniques previously described. First, we make a statistical study of the different quantities used as components of the vector of characteristics for the classes C and I in the training set. Every parameter is described in terms of their mean and variance. The results are shown in Table 1.

We can see that components 6 and 7 are the most discriminative characteristics. They are related to the number of phonemes that coincide with the word and the sequence of phonemes.

In order to perform some experiments with the test database, we have defined three different tasks, denoted by T1, T2 and T3. The set of keywords considered in the three tasks are not disjoint. Both T1 and T2 have 11 different keywords and T3 has 13. The words out of vocabulary have been modelled using the phonetic models as filler models. We have performed experiments with two different weights for the gramatical transitions, 1.8 and 1.6.

For each of the different tasks a comparison is made of the results obtained without the classifier (NC), the ones with classifier with (LVQ) and without (NLVQ) adaptative learning.

Table 2 shows the initial conditions for the different tasks, that is, the results obtained without the classification process. It is important to note that the percentage of recognition (CP) means the ratio of the number of correct words to the total number of words that appear in the transcription

	1.8		1.6	
Task	CP	Fa/kw/h	CP	Fa/kw/h
T1	91.24	25.19	88.97	18.26
T2	90.30	21.65	87.46	15.99
T3	88.24	17.48	82.85	12.09

Table 2: Percentage of recognition (CP) and Fa/kw/h for NC

	Task	NLVQ		LVQ	
		RR	RI	RR	RI
1.8	T1	17.69	64.28	16.59	66.26
	T2	10.20	55.19	9.96	54.73
	T3	9.90	54.80	11.60	48.36
1.6	T1	16.82	63.85	15.92	67.14
	T2	9.67	54.34	9.55	53.09
	T3	9.35	48.76	10.44	59.50

Table 3: Percentages of the relative reductions of recognition (RR) and of insertions (RI) for the different tasks.

and Fa/kw/h is the number of insertions by keywords and hour.

Table 3 shows the relative reductions of the recognized words (RR) and inserted words (RI) after the two different classification processes. It can be observed that the reduction in the inserted words is significantly greater than the reduction in the recognized words, indicating the quality of the two classification processes.

Table 4 shows the averaged cross entropy. We can see an increase in this parameter for every task.

Finally, we observe that the effects of the classifier are similar for both initial conditions studied.

7. CONCLUSIONS

In this study we have explored a procedure to classify the hypothetical word output from a recognizer as correct or incorrect. The procedure is based on taking advantage of the percentage of phonematic recognition. The results obtained show an improvement in the performance of the recognizer by reducing the number of false alarms with a significant less reduction of the correctly detected words. The main problem observed is for short words with 3 to 5 phonemes which are the words with the highest probability of an incorrect classification. Work is in progress to avoid this by substituting the best sequence of phonemes by a lattice of phonemes with depth 2 in order to increase the number of coincidence phonemes.

	1.8			1.6		
Task	NC	NLVQ	LVQ	NC	NLVQ	LVQ
T1	-0.45	-0.30	-0.27	-0.39	-0.25	-0.22
T2	-0.38	-0.25	-0.25	-0.32	-0.20	-0.20
T3	-0.50	-0.33	-0.27	-0.41	-0.29	-0.25

Table 4: CREP for the different tasks and classifiers

8. REFERENCES

[1] R.C. Rose, Keyword detection in conversational speech, Computer Speech and Language, 309–333 (1995)

[2] A. Rubio, P. García, A. de la Torre, J.C. Segura, J. Diaz, M.C. Benítez, V. Sánchez, A. Peinado J.M. López and J.L. Pérez, STACC: an automatic service for information access using continuous speech recognition through telephone line, EUROSPEECH European Conference on Speech Communication and Technology 1779-1782 (1997)

[3] F. Casacuberta, R. García, J. Llisterri, C. Nadeu, J.M. Pardo, A. Rubio. Developmente of Spanish Corpora for Speech Research (Albayzin). Proc the Worksop on International Cooperation and Standardization of Speech databases and Speech I/O assement methods. Chiavari (1991).

[4] T. Kohonen, The Self–Organizing Map Neural Networks, in Theoretical Foundations and Analysis. Edited by Clifford Lau. IEEE Press

[5] S. Cox and R.C. Rose, Confidence measures for the switchboard database, Proc. IEEE Conf. on Acoustic, Speech and Signal processing 511–514 (1996)

[6] T. Schaff and T. Kemp, Confidence measures for spontaneous speech recognition, Proc. IEEE Conf. on Acoustic, Speech and Signal processing 875–878 (1997)

[7] M. Weintraub, F. Beaufays, Z. Rivling, Y. Konig and A. Stolcke, Neural–networks based measures of confidence for word recognition, Proc. IEEE Conf. on Acoustic, Speech and Signal processing 887–890 (1997)

[8] L. Gillick, Y. Ito and J. Young, A probabilistic approach to confidence estimation and evaluation, Proc. IEEE Conf. on Acoustic, Speech and Signal processing 879–882 (1997)