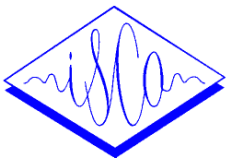


2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

SELECTIVELY TRAINED NEURAL NETWORKS FOR THE DISCRIMINATION OF NORMAL AND LOMBARD SPEECH

Yolande ANGLADE (1) (2), Dominique FOHR (1), Jean-Claude JUNQUA (3)

(1) CRIN-CNRS / INRIA Lorraine, BP 239, 54506 Vandoeuvre-lès-Nancy cédex, FRANCE
(2) SOLLAC, 57191 Florange cédex, FRANCE
(3) SPEECH TECHNOLOGY LABORATORY, Div. of Panasonic Technologies, Inc., 3888 State Street, Santa Barbara, California, 93105, USA
E-mail : anglade@loria.fr, fohr@loria.fr, jcj@crl.mei.co.jp

10.21437/ICSLP.1992-175

ABSTRACT

The purpose of this work is to improve the automatic recognition of confusable words, considering such typical examples as French and American-English Alphabets. Our study proposes a comparison between global methods like DTW or HMM and a new method using neural networks. This method is based on the search for 2 discriminative frames inside the confusable words bearing the distinction between them.. Then a parametrization is done and resulting vectors are given to neural networks. The tests conducted on normal speech, Lombard speech without additive noise and Lombard speech with additive noise show a general improvement of the recognition accuracy.

INTRODUCTION

Performances of global speech recognition methods based on word references, such as the DTW or HMM algorithms, are now quite satisfactory. However, a weak point remains, regarding the recognition of confusable vocabularies. In such cases, discrimination has to focus on the differences between similar words.

By definition, the information useful to make such a discrimination is only present in a small part of the utterance (usually limited to one phoneme). If no discrimination is made, minor differences of longer duration may outweigh the important differences and thus determine the recognition results. It was recently reported that recognition accuracy has been significantly improved for acoustically similar words by using neural network classifiers. Lang and al. obtained a 93% recognition accuracy in multi-speaker experiments on the letters B, D, E, V [1], and Fanty and Cole obtained a 94.2% accuracy in speaker-independent recognition for the same subset [2]. However, much work has still to be done to improve recognition accuracy, especially in adverse conditions when speech is produced in noise.

In this paper, a new discrimination procedure is proposed [3], based on artificial neural networks. First, we are going to describe the characteristics of this method. Then, we will give the recognition results together with a comparison with those given by global methods. These experiments were mainly done on the American-English and French alphabet, since they contain several confusable subsets.

METHOD

1 Databases

1.1 French database

The French database used in our experiments is the BDSONS [4] corpus from the GRECO project. 26 speakers (13 males, 13 females) produced 16 repetitions of the alphabet vocabulary in an isolated-word way. The corpus has been sampled at a 16kHz frequency.

We used this database to compare the DTW algorithm and our new method in speaker-dependent experiments.

Using global methods to recognize the letters of the alphabet led to the following main errors : confusions between A and K; P and T; U and Q; B, D, and V; L, M and N. Consequently, we used these five confusable subsets for our study. They will be noted as $(A,K)_f$, $(P,T)_f$, $(U,Q)_f$, $(B,D,V)_f$, $(L,M,N)_f$ in this article (*).

(*) : index $_f$ will be used in the text to refer to the French subsets

1.2 American-English database

We used an alphabetic vocabulary produced both in normal and noisy conditions (two repetitions for each condition) by 24 speakers. In all tests 12 speakers were used for the training data and the 12 others for the test data. Two combinations of speakers were explored. To simulate speech production in noisy conditions, white-Gaussian noise was injected through headphones at 85dB SPL. To test different types of noise disturbances, some experiments were run for three additive noises extracted mainly from the RSG-10 [5] noise database: white-Gaussian noise, car noise, and babble noise. Several signal-to-noise ratios (SNR) have been considered : no additive noise, 15dB, and 5 dB.

This database allowed us to make speaker-independent comparisons between HMM algorithms and the new method. Results for HMM algorithms have been obtained using manual word boundaries for normal speech and Lombard speech [6] without noise, and automatic word boundaries [7] for Lombard speech with additive noise.

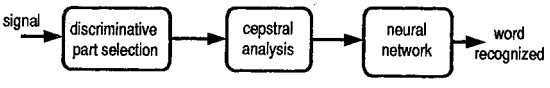
Among this database, 5 confusable subsets have been studied : $(A, J, K)_e$, $(M, N)_e$, $(B, C, D, E, G, P, T, V, Z, three)_e$, $(Go, No, Oh)_e$ and $(F, S, X)_e$. (*)

(*) : index $_e$ will be used in the text to refer to the American-English subsets

2 Discrimination based on neural networks

The main idea of the procedure (see figure 1) is to start the recognition process by locating the discriminative frames of the words. When expert phoneticians have to distinguish letters like P and T, or A and K, they use acoustic cues like vowel formant transitions and burst shape. In these cases, the discriminative information is located at the beginning of the vowel or just before. Our method consists in localizing two frames at these loci, the position of these frames depends on the different subsets of the vocabulary studied. Then, a parametrization is done on these frames and the resulting vectors are given to a neural network.

Figure 1 : block diagram of the method



2.1 Location and extraction of the discriminative frames

Phonetic knowledge

If we consider the different confusable subsets of the databases studied, the discriminative frames are located :

- ☐ for French :
  - in the burst and vowel formant transitions for (P, T) $_f$
  - in the vowel formant transitions and at the place where a burst may be present, for (U, Q) $_f$ and (A, K) $_f$
  - in the final consonant for (L, M, N) $_f$
  - at two locations for (B, D, V) $_f$:
    - + burst and vowel formant transitions for B and D
    - + the fricative part for V.
- ☐ for American-English :
  - in the vowel formant transitions and at the place where a burst may be present, for (A, J, K) $_e$
  - at the place where a possible initial consonant may be detected, for (B, C, D, E, G, P, T, V, Z, three) $_e$ and (Go, No, Oh) $_e$
  - in the final consonant for (M, N) $_e$ and (F, S, X) $_e$.

Extraction method

Phonetic studies have shown the importance of bursts and vowel formant transitions for the recognition process. On the other hand, we know that it is very difficult to detect accurately a burst or to perform formant tracking, and it is even more difficult to do so in the case of noisy speech.

Thus, we detect the vocalic part of the word, and then locate the discriminative frames in the portion of speech located before or after this vowel, where the acoustic cues which we are looking for can be found.

The vocalic part is determined by searching the more energetic frame in the word. Starting from this frame, the algorithm finds a reference point which has its energy 2 or 3 times lower than the more energetic frame. Using a relative energy threshold to locate this point yields the method independent from the speaker loudness. This aspect is very interesting, especially for noisy speech. We have located the discriminative frames relatively to the reference point and performed extensive tests to find their optimal position.

In fact, we found that global energy is an efficient mean to detect the vocalic part of words like P or B. However, when dealing with L, M or N, global energy on the consonant portion may remain at the same level. To solve this problem, we designed a bandpass filter, which allows us to conserve only the energy of the second and third formants, and reject that of the first formant, which appears to approximately be at the same frequency and intensity for both vowel and consonant.

In order to take into account the dynamic aspect of speech, especially in the case of vowel formant transitions, it is interesting to extract at least two discriminative frames. We validated this hypothesis by conducting experimental tests.

2.2 Parametrization

Following the selection of discriminative frames, a 12th order Mel-cepstral parametrization has been done on 51 msec windows. Different numbers of cepstral coefficients, from 8 to 24, have been tested to study the influence of the parametrization order on the recognition accuracy. The results we obtained show that there is no significant change due to this parameter.

2.3 Characteristics of the neural networks

The topology of the back-propagation neural networks is the following : 12 or 24 input neurons (depending on the number of selected frames), 1 hidden layer with 5 neurons and 1 output neuron per word in the subset. Experiments have been conducted using the G^σ tool [10] developped at the CRIN-CNRS & INRIA Lorraine laboratory.

The training was done with 4 repetitions of normal speech for each word and each speaker, in all the tests involving the French speaker-dependent and French speaker-independent database and with 2 repetitions of normal speech for the tests involving the American-English database.

3 Global methods

3.1 Recognition on the French database

We compared the results of the new method to those obtained with the DTW algorithm for speaker-dependent experiments. We used vectorial quantization (256 classes), 12 Mel-cepstral coefficients (same as for neural networks), and an euclidian distance. Four templates in the speaker-dependent mode have been selected as the reference templates, and the remaining 12 others have been used for the tests.

3.2 Recognition on the American-English database

The HMM recognizer used in the evaluation was described in [8]. It is a VQ-based recognizer which uses four types of regression features : R0 with a 10msec window, R1 with a 90 msec window, R1 with a 250 msec window and R2 with a 250msec window. For each type, 12 coefficients are used. The regression features were extracted from the weighted cepstral coefficients derived from the twelfth model order of perceptually-based linear prediction analysis (PLP [9]). For all the tests clean training data (speech produced in a normal environment without background noise) were used.

RESULTS

1 Influence of the location of the discriminative frames

The extraction method described above allows us to locate a reference point in the words of the subsets. To study the influence of the location of the discriminative frames on the recognition rates, we moved a window relatively to this reference point. Figure 2 shows the speaker-independent recognition results obtained for the (P, T) $_f$ subset. On the horizontal axis, the location of the discriminative frame is indicated in msec and takes its origin at the reference point. The vertical lines situated on each point of the curve correspond to the confidence interval of the recognition rates. These intervals, statistically computed, depend on the rate and on the number of values for which it has been obtained. On this figure, we can observe clearly that the discrimination is maximal around the reference point (+/- 8ms). Besides, if the frame is located before the reference point, recognition scores decrease since we extract discriminative frames at the P or T occlusions, which do not contain any useful information. Similarly, if the frame is located after the reference point - i.e. inside the vowel - performances are also very bad.

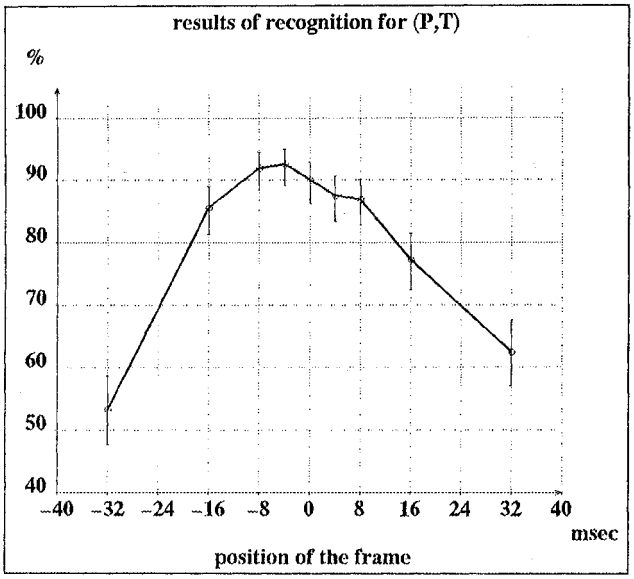


Figure 2 : influence of the location of the discriminative frame.

2 Results on the French database

2.1 Speaker-dependent results

Speaker-dependent results are given in Table 1 for each subset studied . We can notice that the new method based on neural networks gives an important improvement compared to DTW algorithm. Global recognition performances increase from 79.2% to 94.6%. Improvement is obtained for all the subsets, and especially for (P, T) $_f$, (B, D, V) $_f$, and (L, M, N) $_f$. These results show the benefit of only considering the discriminative frames rather than doing a global pattern recognition.

Table 1 : speaker-dependent results obtained by DTW and neural networks on the French database.

FRENCH : SPEAKER - DEPENDENT		
Subsets	ANN	DTW
(P, T) $_f$	97%	80%
(U, Q) $_f$	99%	97%
(A, K) $_f$	99%	93%
(B, D, V) $_f$	88%	75%
(L, M, N) $_f$	93%	62%
TOTAL	95%	79%

2.2 Speaker-independent results

Speaker-independent results obtained by the new method are reported in Table 2. They also validate the new approach in this configuration. The recognition scores are similar to those obtained for speaker-dependent experiments. In the case of the subset (B, D, V) $_f$, we can even observe an improvement of the recognition accuracy, which can be explained by an improvement of the neural networks training.

Table 2 : speaker-independent French results obtained by neural networks.

FRENCH : SPEAKER - INDEPENDENT	
Subset	ANN
(P, T) $_f$	94%
(U, Q) $_f$	99%
(A, K) $_f$	99%
(B, D, V) $_f$	93%
(L, M, N) $_f$	85%

3 Speaker-independent results on the American-English database

3.1 Normal speech

The results obtained on the American-English database for normal speech are reported in Table 3. We can notice that the new method also gives better results than those given by the HMM algorithm for almost all the subsets studied, except for (F, S, X) $_e$ for which similar results are obtained. As a matter of fact, unlike the other subsets for which there is only a one-phoneme difference between the letters, in the subset (F, S, X) $_e$ the letter X [ϵ k s] differs from S and F by two phonemes. This particularity makes the localization of the discriminative frames more difficult.

Table 3 : speaker-independent results obtained for American-English normal speech by HMM and neural networks.

AMERICAN-ENGLISH : NORMAL SPEECH		
Subsets	ANN	HMM
(A, J, K) $_e$	99.3%	95.1%
(B, C, D, E, G, P, T, V, Z, three) $_e$	76.7%	66.9%
(Go, No, Oh) $_e$	100.0%	91.7%
(M, N) $_e$	91.7%	83.3%
(F, S, X) $_e$	88.2%	88.2%
TOTAL	86.3%	79.1%

3.2 Lombard speech

Table 4 shows the results obtained for American-English Lombard speech without any additive noise. As we can see, the recognition difficulty is more important and the recognition rates obtained are lower than that for normal speech. In both cases, ANN and HMM, the spectral modifications due to the Lombard effect, which are not taken into account by the training done with normal speech, have an influence on the recognition performances. However, we can notice that the performances remain better using ANN, except for (F, S, X) $_e$ for which the problems encountered in normal speech stay the same.

If we compare this experiment to the previous one on the same subsets, we can notice that the discriminative frames which give the best results are not the same for normal and Lombard speech. This can be explained by the fact that some spectral characteristics may be more affected by the Lombard effect than others.

Table 5 : speaker-independent recognition percentages obtained forAmerican-English Lombard speech with additive noise by HMM and neuralnetworks.

AMERICAN-ENGLISH : LOMBARD SPEECH WITH ADDITIVE NOISE									
Subsets	method	babble noise		car noise		white-Gaussian noise		TOTAL	
		snr 15 dB	snr 5 dB	snr 15 dB	snr 5 dB	snr 15 dB	snr 5 dB	snr 15 dB	snr 5 dB
(A, J, K) $_e$	ANN	68.7	42.4	72.9	56.2	75.7	69.4	72.4	56.0
	HMM	56.9	38.8	66.7	61.1	58.3	37.5	60.6	45.8
(B, C, D ,E ,G ,P ,T ,V ,Z , Three) $_e$	ANN	31.9	18.3	48.1	39.6	14.1	11.9	31.4	23.3
	HMM	31.2	18.9	40.3	33.3	25.2	10.4	32.2	20.9
(Go, No, Oh) $_e$	ANN	68.8	63.9	71.5	69.4	64.6	61.8	68.3	65.0
	HMM	59.7	52.1	80.6	63.2	72.2	65.2	70.8	60.2
(M, N) $_e$	ANN	84.4	69.9	92.7	92.7	88.5	69.8	88.5	77.5
	HMM	80.2	69.8	82.3	78.1	82.3	66.7	81.6	71.5
(F, S, X) $_e$	ANN	64.6	43.8	72.9	72.2	59.7	46.5	65.7	54.2
	HMM	64.6	45.8	61.8	52.8	58.3	48.6	61.6	49.1
TOTAL	ANN	52.1	36.8	62.8	55.9	43.7	37.7	52.9	43.5
	HMM	48.4	35.1	56.9	48.6	46.8	32.9	50.7	38.9

Table 4 : speaker-independent results obtained for American-English Lombard speech (without additive noise) by HMM and neural networks.

AMERICAN-ENGLISH : LOMBARD SPEECH WITHOUT ADDITIVE NOISE		
Subsets	ANN	HMM
(A, J, K) $_e$	92.4%	79.8%
(B, C, D ,E ,G ,P ,T ,V ,Z , Three) $_e$	55.6%	51.5%
(Go, No, Oh) $_e$	93.1%	91.6%
(M, N) $_e$	87.5%	80.2%
(F, S, X) $_e$	83.3%	88.2%
TOTAL	73.2%	69.3%

3.3 Lombard speech with additive noise

All the results corresponding to the tests conducted on Lombard speech with additive noise are given in Table 5.

Considering this table, we can separate the results for the different subsets as follows : for (A, J, K) $_e$ and (M, N) $_e$, there is a great improvement of the recognition accuracy (more than 10%) using the ANN method. For (F, S, X) $_e$, the two methods give approximately the same results for two kinds of noise (babble and white-Gaussian). It is interesting to notice that it is not the case for Lombard speech without noise for which HMM has better performances. However, for car noise, the improvement given by the new method is important. More generally, results show that recognition is less disturbed by car noise than by the two other noises : indeed, car noise is concentrated on low frequencies (<1000Hz) while white-Gaussian and babble noises are distributed on the whole spectrum. Finally, if we look at (B, C, D, E, G, P, T, V, Z, Three) $_e$ and (Go, No, Oh) $_e$, results are less homogeneous and total percentages are equivalent.

An important point to consider is the speech parametrizations used for the two methods : HMM used 12 static PLP cepstral coefficients and 36 regression coefficients (a total of 48 coefficients), while ANN used only 12 static cepstral coefficients.

CONCLUSION

The new method presented in this paper aims to improve the performances of automatic speech recognition for confusable vocabularies. It is based on the fact that using only discriminative frames of a word can be more efficient than using its global form.

Tests have been conducted on both French and American-English database for normal speech without additive noise. They show that the new method gives equal or better results than DTW or HMM algorithms on all the confusable subsets studied. For the experiments made on Lombard speech without noise, these conclusions remain true, except for the subset (F, S, X) $_e$. Finally, results obtained on Lombard speech with additive noise show an improvement which varies depending on the subset and on the type of noise. An interesting point is the fact that the new method does not require any detection of word boundaries, which are difficult to locate in presence of noise. The new method is based on a reference point taken inside the vowel of each word, which is not too difficult to determine, even in noisy conditions.

Our purpose is now to improve the accuracy of the detection of our reference point, which is an essential aspect of this method, as we have shown in this article. Another future direction of research is the test of other some acoustic front-end using regression coefficients. This could improve the recognition performances for Lombard speech with additive noise.

BIBLIOGRAPHY

[1] K.J. Lang, A.H. Waibel and G.E. Hinton. "A Time-delay Neural Network Architecture for Isolated Word Recognition". vol. 3, No 1, p. 23–43, Neural Networks 1990.

[2] M. Fanty and R. Cole. "Speaker-Independent English Alphabet Recognition : Experiments with the E-Set". p 1361–1364, ICSLP, 1990 Kobe.

[3] Y. Anglade, D. Fohr and J.M. Pierrel. "Reconnaissance de vocabulaires difficiles à l'aide de réseaux neuronaux". p 399–404, JEP, 1992 Bruxelles.

[4] R. Carre, R. Descout, M. Eskenazi, J. Mariani and M. Rossi. "The French Language Database : Defining, Planning, and Recording a Large Database". 42.11, IEEE ICASSP, 1984 San Diego.

[5] H. Steeneken and F. Geurtsen. "Description of the RSG-10 Noise Database". Technical Report, TNO Institute for Perception, 1990.

[6] E. Lombard. "Le signe de l'élévation de la voix". Ann Maladiers Oreille, Larynx, Nez, Pharynx, 37:101–119, 1911.

[7] B. Mak, J.C Junqua, B., B. Rives. "A Robust Speech/Non-Speech Detection Algorithm using Time and Frequency-based Features". I-269 IEEE ICASSP, 1992 San Francisco

[8] T. Applebaum and B. Hanson. "Robust Speaker-Independent Word Recognition using Spectral Smoothing and Temporal Derivatives". p 1183–1186, EUSIPCO, 1990 Barcelona.

[9] H. Hermansky, B. Hanson and H. Wakita. "Low-Dimensional Representation of Vowels Based on All-Pole Modelling in the Psychophysical Domain". Speech Communication, (4):181–187, 1985.

[10] Y. Gong and J-P. Haton. "Towards a General Signal Interpretation System — signal-to-symbol conversion level". p 79–84, IEEE ICPR, 1990 Atlantic City.