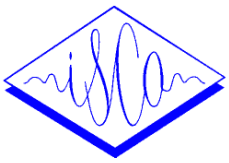

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

ON THE SPEECH FEATURE SELECTION PROBLEM: ARE DYNAMIC FEATURES MORE IMPORTANT THAN THE STATIC ONES?

Jan Nouza

Department of Electrical Engineering,
Technical University of Liberec
Halkova 6, 461 17 Liberec 1, Czech republic
e-mail: jan.nouza@vslib.cz

10.21437/Eurospeech.1995-213

ABSTRACT

The problem of the optimal speech feature selection with respect to CDHMM discrete-utterance recognition is addressed in the paper. Two sequential search algorithms - the Sequential Forward Search (SFS) and Sequential Floating Forward Search (SFFS) have been investigated and applied to several speech databases. It was observed that the dynamic (delta) parameters derived from frame energy and cepstrum were identified as more important than the static ones. The hypothesis about the high discriminative power of the dynamic features has been confirmed in a series of speaker-independent tests. We have even demonstrated that the recognition rate achieved with only the dynamic features closely approached that of the complete feature set. As a practical application of the study, we have proposed a two-level HMM classification scheme that may significantly reduce recognition time without a loss of accuracy.

1. INTRODUCTION

Nowadays, similarly like 10 years ago, spectral (FFT-derived or cepstrum-based) parameters have been the most widely employed features in speech recognition. What has changed is the size of feature vectors. Originally, a frame vector was composed of 8 to 16 *static* (or instantaneous) spectral parameters. Furui [1], and then some other authors, showed the importance of *dynamic* features for speech perception and recognition. The frame vector has been augmented by the first, and more recently also by the second, time derivatives of the spectral parameters (e.g.[2], [3], [4]). At the same time, dynamic features derived from the frame log energy have been also added. Thus, within a short period, the feature vector has trippled its size.

In many reported cases, increasing the number of features really led to improvements in the recognition accuracy, though sometimes the effect was only negligible. The impact on the computational load, however, was evident and undesirable. It became clear that the trend of a mechanical addition of new features should be subject to a critical revision.

Some researchers have used discriminative analysis in search for optimised parameter sets, applicable either for DTW (Dynamic Time Warping) or HMM (Hidden Markov Model) systems (e.g. [5], [6]). An original approach to the feature selection problem, underlaid by large investigations, was presented in Bocchieri's and Wilpon's work [6]. Their method is based on ordering components of a given feature set according to their contributions to the distance measure computed within the HMM evaluation process. The features with the higher contributions are ranked higher and are considered more important for recognition. In several practical vector made only of the high-rank features without a significant loss of recognition accuracy.

In our work, we have investigated another approach to the feature selection problem. We tried to utilize some of the search methods known in the pattern recognition area and apply them in a continuous density HMM (CDHMM) classifier. Unlike the above mentioned rank ordering method, our approach (described in section 2) is based on evaluating features and feature subsets entirely be means of the recognition rate (RR) criterion. Using this criterion, rather than other separability measures, we believe to meet the classifier-specific demands better. In this way, we may cover even such phenomena, like existing correlations between individual features as well as the complex and highly non-linear internal structure of the HMMs.

The most relevant results from experiments performed on four different databases are presented in section 3. In the experiments, discrete-utterance recognition (DUR) tests have been used to order a given set of 18 selected features according to their importance for the CDHMM classifier. The results show that the dynamic features might have significantly higher importance for the recognition compared with the static ones. This, quite suprising, fact has been confirmed in a series of speaker-independent DUR tests. Consequently, the results of the study have been practically applied in a design of a two-level classification scheme described in section 4.

2. SEQUENTIAL SEARCH METHODS

From a large family of feature selection techniques known in the pattern recognition area, we have chosen two for our investigation. Both of them have a sequential character and employ the RR criterion. Their application in an HMM system differs from the common usage just in the fact that the system must be trained first before the recognition rate is evaluated. Obviously, the features used for the parametrization of the testing and the training material must be same.

The Sequential Forward Search algorithm (SFS, see, for example, [7]) orders a given set of P features in the following way: In the first pass, the feature with the highest individual recognition rate is found and selected. In the p-th pass, the set of the p-1 already selected features is augmented always by one of the remaining features and for these p-dimension sets RR values are estimated. The feature which contributed to the highest RR is subsequently included to the set. This is repeated untill all the features are selected. The algorithm can be formally described as follows:

SFS algorithm

1. (Start)

$\boldsymbol{P} = \{x_1 \ldots x_P\}$; // set of P available features

$p = 0$; // number of selected features

$\boldsymbol{S}_p = \varnothing$; // ordered set of selected features

$\boldsymbol{Q}_p = \boldsymbol{P} - \boldsymbol{S}_p$; // set of remaining features

2. (Add a new feature)

$q = P - p$; $p = p+1$;

Do for $i = 1 \ldots q$:

$x_i \in \boldsymbol{Q}_{p-1}$; $\boldsymbol{T}_i = \boldsymbol{S}_{p-1} + \{x_i\}$;

Train and test on feature set $\boldsymbol{T}_i$;

$R(\boldsymbol{T}_i)$ is recognition rate achieved;

$j = \arg\max_{i=1 \ldots q} R(\boldsymbol{T}_i)$; // best of remaining features

$\boldsymbol{S}_p = \boldsymbol{S}_{p-1} + \{x_j\}$; $\boldsymbol{Q}_p = \boldsymbol{Q}_{p-1} - \{x_j\}$;

3. (Stop)

if $p < P$ go to *2.*;

$\boldsymbol{S}_p$ is the requested ordered feature set;

The above presented algorithm may suffer from a nesting effect. It means that the already selected features determine - more or less - the further course of the selection. This could be avoided by search methods that allow also a feature removal, like it is in a family of so called *plus l - take away r* methods. However, a more generalized approach is a floating scheme, represented by the Sequential Floating Forward Search (SFFS, [8]) method. In principle, the SFFS is an SFS procedure provided with a possibility to remove, after each feature addition, that (those) already selected feature(s) that contribute(s) to the achieved recognition rate less than the most recently added feature(s). The number of possible removals is automatically controlled again through the RR criterion. The SFFS method can be described algorithmically in a similar way to the previous method:

SFFS algorithm

1. (Start)

$\boldsymbol{P} = \{x_1 \ldots x_P\}$; // set of P available features

$p = 0$; $\boldsymbol{S}_p = \varnothing$; $\boldsymbol{Q}_p = \boldsymbol{P} - \boldsymbol{S}_p$;

2. (Add a new feature)

$q = P - p$; $p = p+1$;

Do for $i = 1 \ldots q$:

$x_i \in \boldsymbol{Q}_{p-1}$; $\boldsymbol{T}_i = \boldsymbol{S}_{p-1} + \{x_i\}$; estimate $R(\boldsymbol{T}_i)$;

$j = \arg\max_{i=1 \ldots q} R(\boldsymbol{T}_i)$; // best of remaining features

$\boldsymbol{S}_p = \boldsymbol{S}_{p-1} + \{x_j\}$; $\boldsymbol{Q}_p = \boldsymbol{Q}_{p-1} - \{x_j\}$;

3. (Possibly remove the worst feature(s))

Do for $i = 1 \ldots p$:

$x_i \in \boldsymbol{S}_p$; $\boldsymbol{T}_i = \boldsymbol{S}_p - \{x_i\}$; estimate $R(\boldsymbol{T}_i)$;

$j = \arg\max_{i=1 \ldots p} R(\boldsymbol{T}_i)$; // worst of selected features

if $R(\boldsymbol{T}_j) > R(\boldsymbol{S}_{p-1})$ then

$\boldsymbol{S}_{p-1} = \boldsymbol{T}_j$; $\boldsymbol{Q}_{p-1} = \boldsymbol{P} - \boldsymbol{S}_{p-1}$; $p = p-1$; go to *3.*;

4. (Stop)

if $p < P$ go to *2.*;

$\boldsymbol{S}_1 \ldots \boldsymbol{S}_P$ are best selected p-feature subsets;

Inspite of their heuristic character, both the SFS and SFFS do their job well, saving a lot of time that would be otherwise needed for an exhaustive search. Generally, the latter method brings slightly better results as shown in [8], which is obviously paid by greater computation demands. While in the SFS method, the total number of $P(P+1)/2$ feature sets is evaluated, in the latter method, the number is varying with minimum being $P(P+1)$ evaluations. It should be also noted that both the methods are applicable also in cases when the RR values advance in an nonmonotonic way.

3. EXPERIMENTAL WORKS

All the further described experiments have been carried out with the same basic feature set, using a CDHMM training and testing system and employing four different speech databases.

3.1 The basic feature set

The set we have investigated was composed of the following **18 features**: 8 cepstrum coefficients ($c1 \ldots c8$) together with their first derivatives - delta-cepstrum ($dc1 \ldots dc8$) and two dynamic parameters related to the frame energy: delta-energy (*de*) and delta-delta energy (*dde*). The log energy and the lpc-derived and liftered cepstrum were evaluated every 10 ms within 20 ms long frames. A delta-parameter dx at time instant T was estimated from $2K+1$ values of parameter x using the well-known regression formula:

$$dx(T) = \Big(\sum_{t=-K}^{K} x(T+t) \cdot t\Big) \Big/ \Big(\sum_{t=-K}^{K} t^2\Big) \qquad (1)$$

In (1) the K was set to 4, which was a value found optimal in some preliminary experiments. The delta-delta parameter was defined as follows:

$$ddx(T) = dx(T+1) - dx(T-1) \qquad (2)$$

The reason why we did not incorporated delta-delta cepstrum into the studied feature vector followed the practical results presented in [6], where it was shown that the 2nd order derivatives (except of the delta-delta energy) did not play so important role in HMM recognition. One should also note that our basic feature set was nearly identical with that denoted in [6] as an 18-component $DDCEP^+$ set.

3.2 The recognition system

The system employed in the experiments for the feature evaluation was a **CDHMM training/testing system** implemented on a personal computer. Simple left-to-right models with no skips were used to model whole words or whole utterances. The number of model states was set individually for each of the further mentioned databases, but within the databases the models had equal state numbers (see Table 2). State output functions were modelled as mono-mixture normal distributions with diagonal covariances. The Baum-Welch reestimation algorithm was used for training, the Viterbi algorithm for testing.

3.3 Speech databases

The same experiments were always repeated on four different speech databases. All of them consisted of words or multi-word **utterances spoken in isolated way**. Three databases were Czech (CZ) while the remaining was Danish (DK, by courtesy of Aalborg University). Their main characteristics were as follows:

DIGITS (CZ): 10 digits + 6 control words spoken by 8 male and 8 female speakers in 10 repetitions - total number: 2560 tokens.

CAD (DK): 20 highly confusable words (like *bue - buen, kugle - kuglen*, etc.) spoken by 16 male speakers mostly in three repetitions - total number: 930 tokens.

Table 1 - Speech features ordered by the SFS algorithm
(The numbers in each row are feature ordinals estimated for individual databases.)

Feature: Database:	de	dde	c1	c2	c3	c4	c5	c6	c7	c8	dc1	dc2	dc3	dc4	dc5	dc6	dc7	dc8
DIGITS	1	12	16	10	17	7	18	15	11	14	2	3	6	4	5	9	13	8
CAD	1	9	16	10	13	6	17	15	18	3	11	4	2	14	8	5	12	7
DRAW	1	6	12	13	16	17	14	10	18	8	2	4	3	5	7	11	9	15
BUS	1	10	3	16	14	5	13	15	8	18	11	4	2	7	6	9	12	17

DRAW (CZ): 33 control words used in a voice controlled drawing system, 24 male and 24 female speakers of different age, 2 or 3 repetitions per speaker, total number: 4257 tokens.
BUS (CZ): 121 single- and multi-word application-oriented utterances (hours, days, city names, etc.) spoken by 12 male and 12 female speakers via a telephone set, 2 repetitions per speaker, total number: 4840 tokens.

3.4 Experiment conditions

All experiments simulated **speaker-independent tests**. A given database was split into two disjunct parts, from which one was used for training HMMs and the other was employed in testing. Both the testing and training parts were balanced with respect to speakers' gender. To make the results more significant, we repeated each test several times (2 - 4 times) with different database splitting conditions. The rule was that each database item was used once as a test token.

3.5 Experiment results

In preliminary tests we focused on comparing the performance of the SFS and SFFS methods. Practical results showed that the two ordered feature lists did not differ each from other significantly. (It happened only sometimes that the backward steps in the SFFS algorithm changed the order estimated in the forward pass.) A practical drawback of the SFFS method was its high time consumption, about three times higher compared with that of the SFS.

In general, the RR values computed for the optimised subsets selected by the SFS/SFFS algorithms approached very quickly the saturation level (that corresponding to the RR of the complete feature set). In all investigated cases, this level was reached with 10 or even less features. This is demonstrated in Fig.1. Two plots in Fig.1 enable us to compare the RR values belonging to the feature subsets estimated by the SFFS with those found for the subsets created according to the rank ordered list [6]. The difference, which is considerable, particularly, for smaller feature numbers, have two major

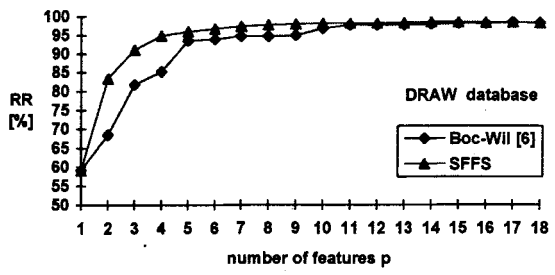

Figure 1- Recognition rates achieved with p-feature subsets created according to the rank-ordering method [6] and the SFFS method.

reasons: The first is obvious - while the list in [6] was created under different conditions and on a different database, the SFFS list is associated with the given database. The other reason reflects the main advantage of the SFFS (or the SFS), which is the RR criterion used for the evaluation. It is evident, for example, that during the feature selection process the SFFS/SFS methods take into account the correlations between the features, while the method introduced in [6] does not do it, at least not in an explicit way.

In Table 1 we present complete ordered lists created for the four databases by the SFS algorithm. Though the absolute feature orders differ between the databases, certain common trends are apparent. We may notice that the delta-energy was always selected as the most important feature. This is no surprise because the same fact was reported also in [6]. However, further observations differ from that of [6]. It was the delta-cepstrum coefficients that were chosen most frequently among the first 10 candidates. The delta-delta energy parameter has found its adequate place in the middle of the lists, the reason being its high correlation with the delta-energy. In general, the static cepstrum coefficients have been identified as the less important features.

In order to verify the hypothesis about the high importance of the delta parameters, we run a series of experiments, the results of which are presented in Table 2. We may observe that the recognition rates achieved with the 8 delta-cepstrum coefficients are unambiguously higher (by 4 to 5%) compared with those achieved with the 8 cepstrum coefficients. Adding the delta-energy and subsequently also the delta-delta energy to the delta cepstrum improved the scores so that they closely approached (in one case even got over) those of the baseline (18-component) feature set.

4. PRACTICAL APPLICATIONS

The above discussed results offer several oportunities for practical applications. The most straightforward one consists in a possibility to reduce a frame vector by removing the less important components, for example, the static ones. As shown in section 3, the impact on the recognition accuracy might be only negligible, while, on the other side, computation time will be reduced by great deal. This could be a useful hint for a design of a simple real-time discrete-utterance recognition system.

We have also proposed a two-level classification scheme based on the feature reduction technique. The scheme combines a *fast match* done with simplified models and an *accurate final match* limited to a small number of preselected standard models. Thus, two sets of HMMs must be available for recognition. In our version, the simplified models have got reduced numbers of states, mixtures and feature components. The set of the accurate models is trained with optimal parameters: the number of states S_A, number of mixtures M_A and number of features P_A. The models for the fast match have

Table 2 - Results of speaker-independent tests performed with different feature sets and different databases

Database:	DIGITS (CZ)	CAD (DK)	DRAW (CZ)	BUS (CZ)
Words/Speakers/Tokens	16/16/2560	20/16/930	33/48/4500	121/20/4840
CDHMM states/mixtures	8/1	10/1	8/1	14/1
Recognition rates [%] achieved with				
8 cepstrum features	90.23	85.05	91.47	89.96
8 delta-cepstrum features	95.31	89.35	96.64	94.17
9 features (de + 8 dcep)	97.58	92.37	97.84	95.91
10 features (de + dde + 8 dcep)	97.76	93.12	97.96	96.28
18 features (complete set)	98.20	92.29	98.36	96.61

parameters: $S_F < S_A$, $M_F < M_A$ and $P_F < P_A$ (where the set of P_F features has been selected according to the above described feature search techniques). The total time needed for the classification of an utterance is given by:

$$T = f(N, S_F, M_F, P_F) + f(N_A, S_A, M_A, P_A) \qquad (3)$$

where f is a function of system parameters, N and N_A are the total number of models and the number of models selected for the accurate match, respectively. The choice of the N_A is constrained by the request that the recognition rate of the two-level scheme must not be worse than that of the standard one-level system. With properly chosen parameters S_F, M_F, P_F a considerable time yield can be achieved.

The described scheme has been used in the design of a practical speech dialogue system ([9]). The system offering information about bus departures operates with the already mentioned BUS database. In the systems's speech recognition unit, the baseline conditions were as follows: $N = 121$, $S = 14$, $M = 1$ and $P = 18$. Using these values for setting the parameters of the accurate models and applying the two-level scheme with parameters $N_A = 20$, $S_F = 4$, $M_F = 1$ and $P_F = 6$ (features *de* + *dc*1...*dc*5), we have reduced the classification time to 28% of its original value, without any loss of recognition accuracy. Now, the complete information system is capable of operating under real-time conditions on a standard personal computer.

5. CONCLUSIONS

In this paper, we want to demonstrate that the problem of speech feature selection is worth of investigating. Since the commonly used feature vectors often contain components with different discriminative power and different level of mutual correlation, feature selection/extraction methods may help in an effective design of a speech recognition system.

We have focused our attention on the investigation of two feature selection methods, the Sequential Forward Search (SFS) and the Sequential Floating Forward Search (SFFS). Their advantage can be seen in the fact that they rely essentially on the recognition rate criterion and thus take into account both the probabilistic characteristics of the features as well as the specific properties of the target classifier. We have demonstrated their practical use in context of the CDHMM discrete-utterance recognition technique.

Analysing the results of experiments conducted on several speech databases, we have arrived at the conclusion that the time-derivative (or delta) parameters might have higher importance for recognition rather than the static (or instantaneous) ones. We explain it by an additional, dynamic, information carried by the time-derivative parameters. The hypothesis has been subsequently verified in a series of speaker-independent discrete-utterance recognition tests. It was observed that the recognition rates obtained with 8 delta-cepstrum coefficients were unambiguously higher (by 4 to 5%) compared with those obtained with 8 cepstrum coefficients. Moreover, a 10-component feature vector composed of delta cepstrum, delta-energy and delta-delta energy manifested nearly the same performance like a complete 18-component vector containing in addition the static cepstral coefficients. It seems that the dynamic features may play a more important, not just a supplementary, role in speech recognition.

Applying the conclusions of the study in practical research, we have proposed a two-level HMM classification scheme that is based on the feature vector reduction. The scheme, eligible for middle-size vocabulary DUR systems, brings a significant performance acceleration without a loss in recognition accuracy.

REFERENCES:

[1] Furui, S.: *Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum.* IEEE Trans. on Acoustics, Speech and Signal Processing. Vol. ASSP - 34, No.1, Feb 1986, pp.52-59.

[2] Hanson, B.A., Applebaum, T.H.: *Robust Speaker-Independent Word Recognition Using Static, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech.* Proc.of ICASSP 90, pp.857-860.

[3] Huang, X.D., Lee, K.F., Hon, H.W., Hwang, M.Y.: *Improved Acoustic Modelling with the SPHINX Speech Recognition System.* Proc.of ICASSP 91, pp.345-348.

[4] Wilpon, J.G., Lee, C.-H., Rabiner, L.R.: *Improvements in Connected Digit Recognition Using Higher Order Spectral and Energy Features.* Proc.of ICASSP 91, pp.349-352.

[5] Lleida, E., Nadeu, C.: *Principal and Discriminant Component Analysis for Feature Selection in Isolated Word Recognition.* Signal Processing V: Theories and Applications. Elsevier Sc. Publ. B.V., 1990, pp. 1251-1254

[6] Bocchieri, E.L., Wilpon, J.G.: *Discriminative feature selection for speech recognition.* Computer, Speech and Language, (1993) 7, pp.229-246

[7] Devijver, P.A., Kittler, J.: *Pattern Recognition: A Statistical Approach.* Prentice-Hall, 1982.

[8] Pudil, P., Ferri, F.J., Novovicova, J. & Kittler, J.: *Floating Search Methods for Feature Selection with Nonmonotonic Criterion Functions.* Proc. of 12th IAPR Conference on Pattern Recognition, Israel, Jerusalem, 1994, pp.279-283.

[9] Nouza, J., Hajek, D.: *A Bus Time-Table Information System with Voice Input and Output.* Proc. of Int. ECMS workshop, Liberec, Czech republic, June 1995, pp.62-65