On The Use Of Principal Component Analysis Over Mel Cepstral Coefficients

Carlos Alberto Ynoguti[1] & Fábio Violaro[2]

[1] INATEL
CP 05, CEP: 37540-000, Santa Rita do Sapucaí, MG
ynoguti@inatel.br

[2]DECOM-FEEC-UNICAMP
CP 6101, CEP: 13083-970, Campinas, SP
fabio@decom.fee.unicamp.br

Abstract - **The purpose of this work is to investigate the application of the Principal Component Analysis (PCA) in the reduction of the dimension of parameter vectors of a continuous speech recognition system. Reduced order parameters should lead to memory and CPU time economy. Preliminary tests in a continuous HMM based system achieved a size reduction of 30% in the parameter vector size without significant performance loss.**

I. INTRODUCTION

Most of speech recognition systems use mel cepstral coefficients with their first and second derivatives as speech parameters. For example, using 12 mel cepstral coefficients, the delta and delta-delta parameters will also be of dimension 12. In this way, for each frame, and for each state in the search space, we have to calculate 36*n* unidimensional Gaussian pdfs (considering a diagonal covariance matrix), where *n* is the number of gaussians in each mixture. For a large vocabulary system (tens of thousands words), the search space will be formed by several thousands of states, and the number of multidimensional gaussian pdfs to be calculated become astronomically high.

Moreover, the calculation of each multidimensional gaussian pdf is relatively complex, as shown below

$$f_{X_1\cdots X_n}(x_1\cdots x_n)=\frac{1}{(2\pi)^{k/2}|\mathbf{C}|^{1/2}}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{X})^t\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{X})} \quad (1)$$

where:
- **C** is the covariance matrix
- $\mathbf{m}_x$ is the mean vector.

A reduction on the dimension of parameter vectors leads to a reduction in computational costs because we have less gaussians to calculate. Memory requirements are also reduced because the number of mean and variances for each multidimensional gaussian is lowered.

With this goal in mind, we adopted the following strategy: first the mel cepstral parameters were grouped together with delta and delta-delta parameters, forming only one parameter vector. With this procedure, the system will work with only one parameter but with dimension 36 instead of 3 parameters with dimension 12. Obviously, this procedure does not reduce the computational load, since we still have to calculate 36*n* gaussians per frame and per state.

Now applying the Principal Component Analysis to the composed vector, one can reduce its dimension with little loss of information. Preliminary tests showed that it's possible to achieve a 30% reduction in vector's dimension without increasing the word error rate.

II. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis is a statistical tool used in a widely range of areas (economy, biology, engineering, etc.) for statistical analysis of multivariate phenomena [3]. It is concerned with explaining the variance-covariance structure through a few linear combinations of the original variables. Its general objectives are data reduction and interpretation. Next, it will be given a brief outline of its principles.

Although *p* components are required to produce the total system variability, often much of this variability can be accounted for by a small number *k*, of the principal components. If so, there is (almost) as much information in the *k* components as there is in the original *p* variables. The *k* principal components can then replace the original *p* variables, and the original data set, consisting of *n* measurements on *p* variables, is reduced to one consisting of *n* measurements on *k* principal components.

Algebraically, principal components are particular linear combinations of the *p* random variables $X_1, X_2, \ldots, X_p$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $X_1, X_2, \ldots, X_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure. The theory of principal component analysis can be resumed in the next two results.

Result 1. Let **S** be the covariance matrix associated with the random vector $\mathbf{X}^t=\left[X_1, X_2, \ldots, X_p\right]$. Let **S** have the eigenvalue-eigenvector pairs $(l_1,\mathbf{e}_1),(l_2,\mathbf{e}_2),\ldots,(l_p,\mathbf{e}_p)$, where $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$. The *i*-th principal component is given by

$$Y_i=\mathbf{e}_i^t\mathbf{X}=e_{i1}X_1+\cdots+e_{p1}X_p, i=1,\ldots,p \quad (2)$$

With this choices,

$$\begin{aligned} \mathrm{Var}(Y_i) &= \mathbf{e}_i^t \mathbf{S} \mathbf{e}_i \\ \mathrm{Covar}(Y_i, Y_j) &= \mathbf{e}_i^t \mathbf{S} \mathbf{e}_j = 0 \end{aligned} \quad (3)$$

This result shows that the principal components are uncorrelated and have variances equal to the eigenvalues of **S**.

Result 2. Let $\mathbf{X}^t = [X_1, X_2, \ldots, X_p]$ have covariance matrix **S**, with eigenvalue-eigenvector pairs $(l_1, \mathbf{e}_1)$, $(l_2, \mathbf{e}_2), \ldots, (l_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
Let $Y_i = \mathbf{e}_i^t \mathbf{X} = e_{i1} X_1 + \cdots + e_{p1} X_p, i = 1, \ldots, p$ be the principal components. Then

$$\begin{aligned} s_{11} + s_{22} + \cdots + s_{pp} &= \sum_{i=1}^{p} \mathrm{Var}(X_i) \\ &= l_1 + l_2 + \cdots + l_p = \sum_{i=1}^{p} \mathrm{Var}(Y_i) \end{aligned} \quad (4)$$

This result shows that the total variance is given by

$$s_{11} + s_{22} + \cdots + s_{pp} = l_1 + l_2 + \cdots + l_p \quad (5)$$

and consequently, the proportion of total variance due to (explained by) the *k*-th principal component is

$$\frac{l_k}{l_1 + l_2 + + l_p}, \quad k = 1,2,\ldots, p \quad (6)$$

If most (for instance, 80% to 90%) of the total population variance, for large *p*, can be attributed to the first one, two or three components, then these components can "replace" the original *p* variables without much loss of information.

Its important to note that although the above results have been derived for the covariance matrix **S**, the same results would be achieved using the correlation matrix **R**.

III. EXPERIMENTAL SETUP

Recognition engine

For the tests, a continuous HMM locally developed software [5] was used. This system uses context independent phones as basic units, the One Pass algorithm [2] as search engine, and mel-cepstral coefficients together with their first and second derivatives.

To find the eigenvalues and eigenvectors of the covariance matrix, we used the QL algorithm with implicit shifts in conjunction to the Householder tridiagonalization procedure [4].

The recognition times were obtained from an AMD-K6II 350 MHz, 64 MB RAM machine running under Windows98® platform.

Database

The sentences were chosen from a work from Alcaim et. al. [1], where 200 phonetically balanced sentences were listed. In these sentences we counted 694 different words.

For the recordings, 40 adult speakers (20 males, 20 females) were selected. The test set consisted of 5 speakers of each gender arbitrarily chosen; the remaining ones formed the training set. In gender dependent tests, the training and test sets were extracted from the previous ones, resulting in 5 test speakers and 15 training speakers each.

An extra male speaker completes the database. He spoke all the 200 sentences four times. Three of them were used to train the system and the last one for the tests. These locutions were used for speaker dependent tests.

All the recordings were performed in an office environment, with a SHURE SM-58® directional microphone, using a SoundBlaster AWE 64® sound card. The sentences were recorded at 11025 kHz sampling rate and 16 bits resolution.

The locutions were manually transcripted, using the Cool Edit 2000® software for viewing the waveform and spectrogram and earphones to listen to them carefully.

IV. TESTS

The tests were performed in three modes:

- speaker dependent,
- gender dependent (only female speakers, only male speakers),
- and speaker independent.

The reasons for these choices were:

1. the great majority of commercial speech recognition systems asks the gender of the user. It's easy to conclude that they don't operate in real speaker independent mode, but in gender dependent mode, therefore facilitating the task of the recognizer;
2. also, most of these systems have an adaptation period, where a new user is invited to train the system so as to track his/her personal features;
3. principal component analysis is closely related to sample space variance, which has direct relation with the number of speakers involved. In other words, it's reasonable to expect that one can achieve a greater compression in speaker dependent mode compared to speaker independent mode, with the same performance loss. Surprisingly, this hypothesis was not confirmed in the our tests. In Section IV we present some hypothesis to explain this behavior.

Initial tests

The first test series were performed using the mel-cepstral, delta mel-cepstral and delta-delta mel-cepstral coefficients as three different parameters. The objective of these initial tests is to provide a reference performance, based on which all the subsequent results will be compared.

Tests were made with 4, 5, 6 and 7 gaussians in the mixture for each state, and the best results, shown in Table 1, were obtained using 5 gaussians per mixture. In

this Table and in the following ones, the symbols D, S and I correspond respectively to deletion, substitution and insertion percentage errors. The recognition time is given in minutes.

Tests with combined parameters

The second set of tests consisted in combining the mel-cepstral, delta and delta-delta parameters in only one parameter vector. Now we have only one vector of dimension 36 instead of three vectors of dimension 12. This procedure does not affect the computational and memory requirements, but it was performed just to see what happens with the word error rate when combining the feature parameters in a single parameter vector.

As in the previous ones, we made tests using 4 to 7 gaussians in the mixture, and the results are shown in Table 2.

Table 1. Initial tests results.

Tests	D (%)	S (%)	I (%)	total (%)	# gauss	time (min)
Dep.	0.84	2.51	2.05	5.40	5	2:44
Male	3.88	8.22	2.21	14.31	5	2:06
Female	2.81	10.58	3.04	16.44	5	3:05
Indep.	3.69	11.68	2.82	18.19	5	2:43

Table 2. Combined parameters tests results.

Tests	D (%)	S (%)	I (%)	total (%)	# gauss	time (min)
Dep.	1.67	2.51	1.07	5.25	4	3:11
Male	2.74	7.53	2.21	12.48	6	2:42
Female	3.04	12.33	3.80	19.18	5	2:56
Indep.	3.92	12.79	2.32	19.03	5	2:53

Observing Table 1 and Table 2, we observe that the tests with a single speaker and with male speakers achieved a slightly better performance by grouping all three parameters in a single one, while the speaker independent and female speaker tests presented a little performance loss. In general, we can conclude that there are no great changes neither in the word error rate nor in the number of gaussians when combining the parameters in a single vector.

Tests with principal component analysis

The final set of tests were performed using the principal component analysis over the combined vectors to reduce their dimension.

The first question to be answered is: how much compression can be achieved without a significant performance degradation? In Section II we showed that we can use (5) to estimate the information loss with the number of principal components. Taking the locutions of the speaker independent training set, we obtain the results shown in Table 3 for the correlation matrix and Table 4, for the covariance matrix.

Repeating this tests with all the other sets (male speakers, female speakers and speaker independent), similar results were obtained. This behavior contrasts our expectation of a possibly higher compression rate with less inter-speaker variability, pointed out on item c) of Section IV.

Observing Table 3 and Table 4 we can conclude the following:

- there is no substantial differences in using the correlation or the covariance matrix.
- a reduction from 36 original variables to 20 principal components keeps about 99% of the total information, which seems to be a reasonable reduction threshold, i.e., there would be a marginal loss in the word error rate by substituting the 36 original variables with the 20 principal components.

Table 3. Proportion of total variance versus number of principal components. Results derived from correlation matrix for speaker dependent training set.

# comp.	% variance	# comp.	% variance
1	39.95%	19	98.89%
2	65.29%	20	99.04%
3	73.43%	21	99.17%
4	80.54%	22	99.30%
5	84.14%	23	99.41%
6	87.17%	24	99.51%
7	89.35%	25	99.60%
8	91.23%	26	99.67%
9	92.71%	27	99.74%
10	93.96%	28	99.80%
11	95.21%	29	99.84%
12	96.16%	30	99.87%
13	96.93%	31	99.90%
14	97.60%	32	99.93%
15	97.95%	33	99.95%
16	98.28%	34	99.97%
17	98.49%	35	99.99%
18	98.70%	36	100.00%

Table 4. Proportion of total variance versus number of principal components. Results derived from covariance matrix for speaker dependent training set.

# comp.	% variance	# comp.	% variance
1	41.40%	19	98.74%
2	61.85%	20	98.91%
3	71.13%	21	99.06%
4	78.45%	22	99.20%
5	82.41%	23	99.33%
6	85.51%	24	99.45%
7	87.99%	25	99.55%
8	90.10%	26	99.63%
9	91.74%	27	99.70%
10	93.18%	28	99.77%
11	94.54%	29	99.82%
12	95.62%	30	99.86%
13	96.50%	31	99.89%
14	97.26%	32	99.92%
15	97.67%	33	99.94%
16	98.04%	34	99.97%
17	98.28%	35	99.98%
18	98.52%	36	100.00%

To confirm these conclusions, several tests were performed using the correlation and covariance matrixes

and various levels of compression. In Table 5 and Table 6, speaker independent with 6 gaussians mixtures tests results are shown.

These results confirm the conclusions of Table 1 and Table 2, that there are little differences in choosing the correlation or covariance matrix, and a good reduction threshold is to use 20 principal components to replace the 36 original variables.

An interesting thing to note is that the recognition times were almost not affected by the reduction in parameter vector sizes, as one would expect. It's possibly because there are other factors that have more influence on the system's overall computer load than the calculation of the Gaussian mixtures.

Table 5. System performance versus number of principal components. Speaker independent mode. PCA over correlation matrix.

# components	D (%)	S (%)	I (%)	total (%)	time (min)
15	5.93	17.16	2.74	25.84	2:41
20	5.14	13.43	2.40	20.97	2:41
25	4.37	13.39	2.66	20.43	2:42
30	3.65	11.34	2.70	17.69	2:42
36	3.65	12.18	3.08	18.91	2:42

Table 6. System performance versus number of principal components. Speaker independent mode. PCA over covariance matrix.

# components	D (%)	S (%)	I (%)	total (%)	time (min)
15	5.71	17.46	2.85	26.03	2:43
20	3.81	13.55	1.86	19.22	2:43
25	4.37	13.93	2.51	20.81	2:43
30	4.11	12.48	2.62	19.22	2:42
36	4.30	13.16	2.47	19.94	2:47

According to these results, a final set of tests were performed using 20 principal components for all the test sets, and the results are shown in Table 7 and Table 8.

Table 7. PCA over correlation matrix, 20 principal components, and 6 Gaussians per mixture.

Tests	D (%)	S (%)	I (%)	total (%)	time (min)
Dep.	1.37	2.44	0.68	4.49	2:16
Male	4.72	9.13	1.29	15.44	2:38
Female	3.58	14.15	2.97	20.70	2:51
Indep.	5.14	13.43	2.40	20.97	2:41

Table 8. PCA over covariance matrix, 20 principal components, and 6 Gaussians per mixture.

Tests	D (%)	S (%)	I (%)	total (%)	time (min)
Dep.	1.90	3.04	1.41	6.09	3:14
Male	4.72	8.83	1.21	14.76	2:30
Female	3.12	15.37	3.73	22.22	2:42
Indep.	3.81	13.55	1.86	19.22	2:43

Comparing Table 1 and Table 2 with Table 7 and Table 8 we see that the results are quite similar, and we can conclude that it's possible to reduce the parameter vector size from 36 to 20 without much loss of performance in terms of word error rate. Also, either the correlation or the covariance matrixes can be used for this purpose, with similar results.

V. CONCLUSIONS

In this work we presented a theoretical outline about Principal Component Analysis and a practical implementation over mel-cepstral, delta and delta-delta combined parameter vectors of a continuous speech recognition system based on continuous Hidden Markov Models.

Tests results showed that it's possible to achieve a 30% reduction in parameter vector size without significant performance deterioration, a result that is consistent with theoretical results.

Similar performances were achieved using either the correlation or the covariance matrixes, indicating that scaling between the coefficients is not crucial for the overall word error rate.

Recognition times do not felt down with vector size reduction as expected, and this fact leads us to conclude that parameter vector dimension doesn't have a major influence on the overall computer load. However, the size reduction leads to proportional memory saving, which is always desirable for high computational resources consuming systems.

Acknowledgements. The authors wish to thank FAPESP agency for partial funding (Proc. 01241-2/99).

VI. BIBLIOGRAPHY

[1] Alcaim, A., Solewicz, J. A., Moraes, J. A. Frequência de ocorrência dos fones e lista de frases foneticamente balanceadas no português falado no Rio de Janeiro. *Revista da Sociedade Brasileira de Telecomunicações*, **7**(1):23-41, 1992.

[2] Ney, H. The use of a one-stage dynamic programming algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-32**(2), April, 1984.

[3] Johnson, R. and Wichern, D. W. *Applied Multivariate Statistical Analysis*. New Jersey : Prentice Hall,1998.

[4] Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. *Numerical recipes - the art of cientific computing*. Cambridge : Cambridge University Press,1987.

[5] Ynoguti, C. A. e Violaro, F. Um sistema de reconhecimento de fala contínua baseado em modelos de Markov contínuos. *Anais do XVIII Simpósio Brasileiro de Telecomunicações*, 3 a 6 de setembro de 2000, Gramado, RS.

Carlos Alberto Ynoguti received the B.S. and M.S. degrees in electrical engineering from São Paulo University – USP - in 1991 and 1994, respectively, and the Doctor of Science degree from State University of Campinas (UNICAMP) in 1999. From 1999 to 2000 worked as researcher at UNICAMP, and now is an Associate Professor at National Institute of Telecommunications – INATEL. His interests are in the

fields of Digital Signal Processing, Speech Processing and Neural Networks.

Fábio Violaro received the B.S., M.S. and Doctor of Science degrees in electrical engineering from State University of Campinas (UNICAMP) in 1973, 1975 and 1980, respectively. Presently he is a Titular Professor at UNICAMP and coordinates the Speech Digital Processing Laboratory. His research are in digital processing of speech area: analysis, coding, recognition and synthesis.