SPEAKER ADAPTATION USING EIGENVOICES TECHNIQUE

Liselene de Abreu Borges[1], Miguel Arjona Ramírez[1], Rubem Dutra Ribeiro Fagundes[2]

[1]Dep. Eng. Eletrônica – Escola Politécnica
Caixa Posta 61548, Universidade de São Paulo
CEP 05424-970 São Paulo, SP, Brazil
Tel.: +55-11-818-5606 Fax: +55-11-818-5718
e-mail: liselene@lps.usp.br, miguel@lps.usp.br

[2]Dept. de Eng. Elétrica – Faculdade de Engenharia
Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS
Tel.: +55-51-3320-3540 Fax: +55-51-3320-3625
e-mail: rubemdrf@attglobal.net

ABSTRACT

This paper discusses speech recognition systems (SRS) using speaker adaptation techniques. The most recent speech recognition systems use Hidden Markov Models (HMM). For such systems, the eigenvoices speaker adaptation technique presents the best performance among other techniques usually suggested by researchers. This performance is due mainly to the limited amount of data necessary to perform speaker adaptation. In our experiments we have reached improvements of around 10% in speaker adaptation system compared with the corresponding independent speaker speech recognition system and just using a very small fraction of speakers' data.

1 INTRODUCTION

A speaker dependent speech recognition system (SD) presents best performance because all data available comes from just one speaker, usually the system user. However, it is necessary to utilize a large amount of data from this user, enough to provide a good recognition performance. At this point, whenever the vocabulary increases, the amount of data for training will increase, becoming very difficult to keep the system's performance. The usual way to solve this problem is to train an independent speaker recognition system (SI) using data from several speakers. Nevertheless, the final performance is not very good compared with the dependent speaker system.

The solution is the building of a speaker adaptive system [1], in which an independent speaker system (SI) is created. After that, using a speaker adaptation technique, the speech recognition system dynamically becomes a speaker dependent system (SD).

An adaptation system (Figure 1) is not a fully trained speaker dependent system, but a system with most of the knowledge taken from an SI system and specific set of information from the new user, extracted from the user's adaptation data.

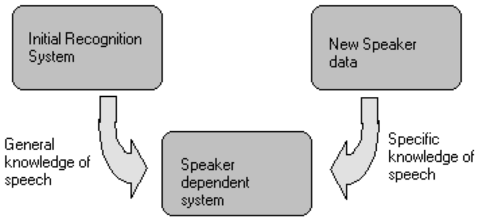

Initial Recognition System
New Speaker data
General knowledge of speech
Speaker dependent system
Specific knowledge of speech

Figure 1 – Speaker adaptation system retains general and specific knowledge of speech

The use of adaptive systems allows improvements in speech recognition performance at low cost. The main goal is to get a better performance from a small set of data and save a lot of computation time.

2 SPEAKER ADAPTATION SYSTEM

A speaker adaptation system (SAS) will try to modify an SI system, previously trained as a speaker dependent system - close to an ideal SD system, for a given new speaker and using just a small set of adaptation data[1].

The process whose adaptation data set is given a priori is usually called supervised adaptation. When the adaptation data set is unknown a priori, the process is called unsupervised adaptation. Furthermore, the adaptation process can be executed directly on the input signal, usually called spectral mapping adaptation [2], or on the HMM parameters which we call model mapping adaptation[3].

Finally, the adaptation process can be defined as *offline* [2] when it runs before the new user can utilize the SRS for the first time[1], or *online* [2] when it runs together with the new speaker for the first time.

2.1 Eigenvoices

The eigenvoices technique [4] is based on an image processing technique [5], namely eigenfaces technique, usually applied as an image compression method. The main point is to reduce the dimension of data variables keeping the best data variation on those remaining parameters[6], because most of these parameters have a high correlation with each other.

Westwood in [7] says: “The eigenvoices form a basis of a subspace of the acoustic model space, and are chosen to account for inter-speaker variability.”. For a given set of parameters estimated from different SDs, the Principal Component Analysis (PCA) will define the linear directions along which most of the data variability lies. Such directions are called principal components or eigenvoices.

2.1.1 Eigenspace estimation

The first step in using eigenvoices is to build the eigenspace [8]. In order to do so, it is necessary to train T different SDs , using T different speakers[2]. Each SD has its parameters annexed. In this work, these parameters are the means of the Gaussian output distributions of the HMM, but we can also use the variances of the output distributions, transition matrices or other parameters.

After that, these means from each t of SD model are copied in a vector called supervector with dimension D where D is the total number of adaptation parameters.

The next setp consists in the buildingof a very large matrix **M,** using all T supervectors, with dimensions (DxT) as follow:

$$\mathbf{M} = [\mathbf{p}^{(1)} \quad \mathbf{p}^{(2)} \quad \cdots \quad \mathbf{p}^{(T)}] \qquad (1)$$

where $\mathbf{p}^{(t)}$ is the supervector (Dx1) with all the *t* speaker parameters and t=1,2,...T.

The eigenspace will be extracted from matrix **M** through Principal Component Analysis (PCA) [6] as we can see below:

$$\mathbf{E} = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_D] = PCA(\mathbf{M}) \qquad (2)$$

Where **E** is the eigenspace (eigenvoices space). Each line of **E** matrix $\mathbf{e}_k$, is given by:

$$\mathbf{e}_k{}^T = \left[e_k^{(1)} \quad e_k^{(2)} \quad \cdots \quad e_k^{(j)} \cdots\right] \qquad (3)$$

where j is the state model and $k = 1,2,...,D$.

2.1.2 Number of eigenvoices components

In a SRS, usually a number D of components is necessary in order to reproduce the whole system variability. However, almost the complete system variability is reduced in a small number of K components. In other words, the most of the system’s information is located in a small set of K principal components which can replace and represent the whole set of D components. Thus, the original data set, composed of T observations of D components, will be reduced to a set of T observations of K principal components [9]. The K should have a value smaller than T=rank(**M**), with K<T<<D, and can be defined by many ways[9]. This work uses the percent cumulative variation as seen below:

$$\%VarCum_K = 100.\frac{\sum_{k=1}^{K} d_k}{\sum_{d=1}^{D} d_d} \qquad (4)$$

where d_k is the eigenvalue associated to eigenvector $\mathbf{e}_k$.

This number is the ration of each eigenvalue (associated with each eigenvoice) to the sum of all D eigenvalues of **E**. Usually K can be choose with the percent cumulative variation value around 80 to 90%.

This new eigenspace $\tilde{\mathbf{E}}$ is now given by:

$$\tilde{\mathbf{E}} = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_K] \qquad (5)$$

where those last (D-K) eigenvoices will be ignored.

2.2 Eigenvoices coefficients

The adaptation parameters are given by:

$$\hat{\mathbf{p}} = \tilde{\mathbf{E}}.\mathbf{v} = \sum_{k=1}^{K} v_k \mathbf{e}_k \qquad (6)$$

where v_k are the eigenvoices coefficients to be estimated [8].

[1] In this case the speaker data set has been acquired before using the system as part of a pre-processing step.

[2] These T speakers are the base speakers of the system

2.2.1 Maximum likelihood Eigen-Decomposition

In order to estimate the eigenvoices components is necessary to maximize the likelihood of adaptation data given the HMM model $\hat{\lambda}$ [10], as seen from:

$$\hat{\lambda} = \arg\max_{\lambda\in\Omega} P(\mathbf{O} \mid \lambda) \qquad (7)$$

where **O** is the observation set that is intended to be represented by the adaptation model, and Ω is the set of HMM. This maximization is given by the maximum likelihood estimation decomposition (MLED) [8], using the maximum likelihood estimation algorithm (ML) [11] in order to calculate the Equation (6).

The ML algorithm transforms the function P(**O**|λ) in an auxiliary Baum function $Q(\lambda,\hat{\lambda})$ maximizing this function in relation t $\hat{\lambda}$ as follows [10]:

$$Q(\lambda,\hat{\lambda}) = \sum_{\mathbf{q}\in\mathbf{Q}} P(\mathbf{O},\mathbf{q} \mid \lambda)\log P(\mathbf{O},\mathbf{q} \mid \hat{\lambda}) \qquad (8)$$

where **q**=(q$_1$,...,q$_T$) is the state sequence and **Q** is the set of all possible state sequences. According to [12] the development of expression (8) will be:

$$\sum_t \gamma_t^j \mathbf{e}_k^{(j)T}\mathbf{\Sigma}^{(j)-1}\mathbf{o}_t = \sum_t \gamma_t^{(j)} \sum_{i=1}^{K} v_i \mathbf{e}_i^{(j)T}\mathbf{\Sigma}^{(j)-1}\mathbf{e}_k^{(j)} \qquad (9)$$

where:

$\mathbf{o}_t$ is the observation vector at a given time *t*;

$\mathbf{\Sigma}^{(j)-1}$ is the inverse covariance matrix of state *j*;

$\gamma_t^{(j)}$ is state *j* occupation probability in time t given the observation sequence **O** and HMM λ.

3 METHODOLOGY

All tests in this work have been done by using an isolated speech recognition system with a 20-word vocabulary in English, from the well known TIMIT speech corpus [13]. Each word is a continuous distribution HMM with 6 states, and each state has one output Gaussian distribution with 12 MFCCs[3] and one frame energy coefficient. The system's vocabulary is given in Table 1:

she	Suit	year	to	rag
had	greasy	don't	carry	like
your	Wash	ask	an	that
dark	Water	me	oily	in

Table 1- system vocabulary

The SI system was trained by 20 speakers (7 women, 13 men) achieving 84,33% correct recognition rate in a test set of 15 speakers (6 women, 9 men).

3.1 Results

The first test was the effect of eigenvoices subspace dimension, changing the number of eigenvoices components or, in other words, changing k (Figure 2).

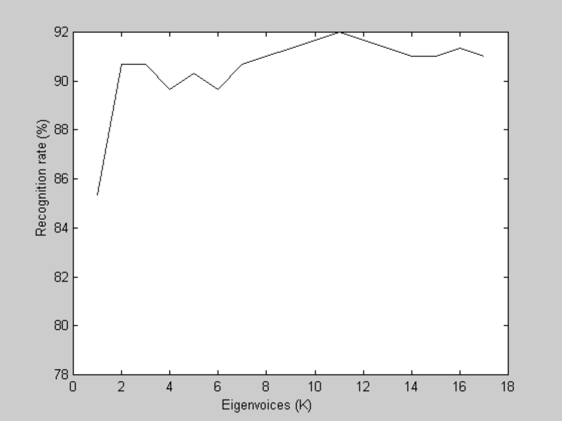


Figure 2- Recognition rate by varying eigenvoices number

Figure 3 shows the percent cumulative variation. We would like to point out that the most percent cumulative variation is concentrated in the first three eigenvoices.

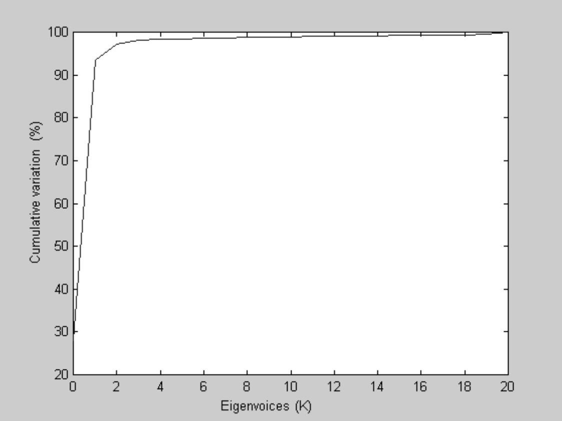


Figure 3 – Eigenvoices cumulative variation

[3] Mel Frequency Cepstral Coefficients

Some tests have been done changing the number of speakers in the SI system. Figure 4 shows that the number of base speakers has no effect in improving marginal system performance. The marginal improvement provided by the adaptation is around an additional 10%.

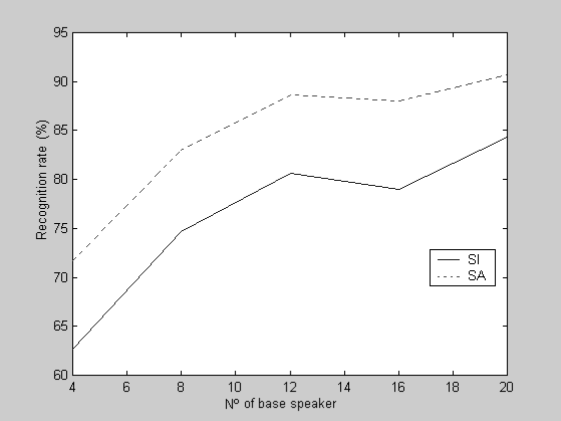
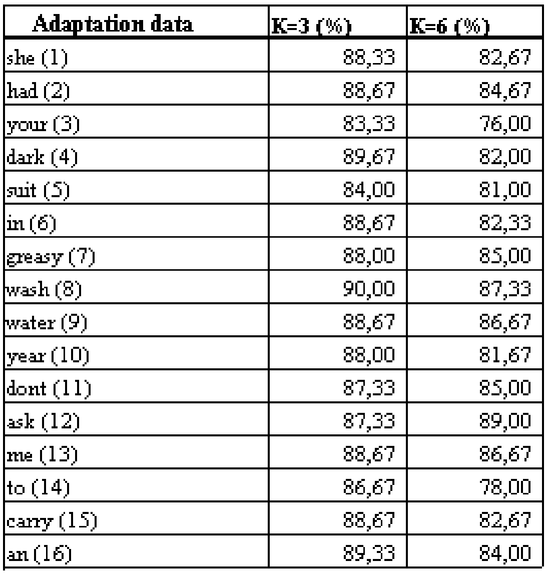


Figure 4 – System independent and system adapted recognition rate for different N^{os} of base speaker of the SI model

Table 2 shows correct recognition rate results, using one word as adaptation data.



Adaptation data	K=3 (%)	K=6 (%)
she (1)	88,33	82,67
had (2)	88,67	84,67
your (3)	83,33	76,00
dark (4)	89,67	82,00
suit (5)	84,00	81,00
in (6)	88,67	82,33
greasy (7)	88,00	85,00
wash (8)	90,00	87,33
water (9)	88,67	86,67
year (10)	88,00	81,67
dont (11)	87,33	85,00
ask (12)	87,33	89,00
me (13)	88,67	86,67
to (14)	86,67	78,00
carry (15)	88,67	82,67
an (16)	89,33	84,00

Table 2 – Eigenvoices adaptation results, having one word as adaptation data, for 1st, 3rd and 6th eigenvoices

The last test was carried out by changing the amount of adaptation data, as shown in Figure 5. This figure demonstrates that the amount of adaptation data has practically no effect in recognition performance. The recognition rate performance using one word as adaptation data was 90%, against 91% recognition rate using all words as the adaptation data set. We can conclude that this method is specially indicated when just a very small amount of adaptation data is available.

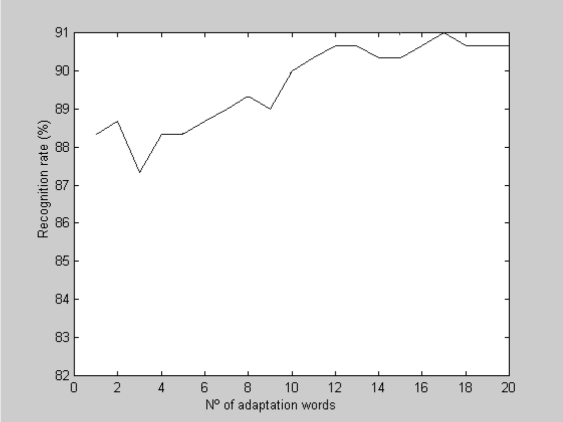


Figure 5 – Speaker adaptation system recognition rate with an increasing size adaptation data set, for first three eigenvoices component

4 CONCLUSIONS

It seems very clear that the dimension K of the eigenvoices plays a main role in the system's behavior, not only due to its influence in system performance, but also given the real advantage in the use of a small fraction of training data to perform speech recognition. In this sense we would like to point out that eigenspace dimension K should be just as small. as the amount of data available to perform speaker adaptation. In the same way, if there is a large amount of adaptation data, K must be properly dimensioned to fit the size of the data set.

It is understandable that the amount of information from the speaker is extracted from the adaptation data and there is a strong relation between thisinformation (in an acoustic sense) and the K dimensions necessary to represent such information in the eigenspace. Accordingly, the first K eigenvoices are chosen from the HMM variability[4] and if we try to use a large K with reduced adaptation data, we would make a bad estimation about these new eigenvoices (or in other words these new dimensions) leading to incorrect parameter estimations.

We would like to stress that using just 70% of the total amount of data the maximum performance was achieved. In most of eigenvoices adaptation system [4] [7] [2] the amount of data came from 100 base speakers. In this work, we have used just 20 base speakers reaching maximum performance.

We also would like to point out that this technique is especially indicated for small vocabulary[5], because a large

[4] And these HMM will be trained from the available data.

[5] There is no system using around 1000 words reported in the literature until now.

one will demand a lot of training time[6]. For large vocabulary SRS, the Maximum Likelihood Linear Regression technique (MLLR) is much more indicated. As a future work we are considering the use of the eigenvoices technique with regression classes [15].

Also in the future, we plan to use the eigenvoices in SRS with acoustic units, like phonemes [16], improving recognition performance.

5 REFERENCES

[1] FURUI, S**.,** ***Speaker-Independent and Speaker-Adaptative Recognition Techniques,*** In: Advances in Speech Signal Processing. Ed. Furui, S., Sondhi, M., New York: Marcel Dekker, pp.597-621,1992.

[2] CHRISTENSEN, H. ***Speaker Adaptation of Hidden Markov Models using Maximum Likelihood Linear Regression***. Thesis, Aalborg University, Denmark, 1996.

[3] WOODLAND, P., ***Speaker Adaptation: Techniques and Challenges,*** Proceedings IEEE Automatic Speech Recognition and Understanding Workshop, pp.85-90, Colorado, 2000.

[4] KUHN, H. ,*et. al*.**,** ***Eigenvoices for speaker adaptation***. Proc. of ICSLP-98, pp.1771-1774, Sydney, Australia, 1998.

[5] KUHN, H. ,*et. al*.**,** ***Eigenfaces and Eigenvoices: Dimensionality reduction for specialized Pattern Recognition***. IEEE workshop on multimedia Signal Processing, California , 1998.

[6] JOLLIFFE, T., ***Principal Component Analysis***, Springer-Veriag, New York, 1986

[7] WESTWOOD, R.**,** ***Speaker Adaptation using Eigenvoices,*** Thesis, Cambridge University, Cambridge, 1999.

[8] KUHN, H. ,*et. al*.**,** ***Eigenvoices for speaker adaptation***. Internal technical report, STL, California, 1997.

[9] JOHNSON, R., WICHERN, D., ***Applied Multivariate statistical analysis***, Prentice Hall, Texas, 1988.

[10] RABINER, L., ***A tutorial on Hidden Markov Models and selected Applications in Speech Recognition***. Proceedings of the IEEE, vol 77. N°2, pp.257-284, 1989.

[11] DELLER, J.; PROAKIS,J.; HANSEN, J. ***Discrete-Time processing of Speech Signals***. Macmillan, New York, pp.63-66, 1993.

[12] BORGES, L., **Speaker adaptation system using eigenvoices**. MSc. Dissertation. In Portuguese. University of São Paulo. São Paulo. 2001.

[13] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST).***The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus***, Virginia, http://www.nist.gov/, 1990.

[14] DAVIS, B.; MERMELSTEIN, P. ***Comparison of parametric representations for monosyllabic word recognition in continuous spoken sentences.*** . IEEE Acoustics, Speech and Signal Proceeding, Vol. 28, pp.357-366, 1980.

[15] LEGGETER, C. ***Improved Acoustic Modelling for HMMs using Linear Transformations***. Ph.D. thesis, Cambridge University, Cambridge, 1995.

[16] FAGUNDES, R. D. R, ***Phonetic-phonologic approach to continuous language speech recognition system.*** Ph.D.thesis. In Portuguese. University of São Paulo – POLI/USP, São Paulo, Brazil, 1998.

[6] This is the case for systems with no phonetic modeling. For large vocabulary SRS with phonetic modeling, the HMM phonetic models can be adapted by eigenvoices technique.