TR-H-296

Channel Distortion Equalization for Robust Speech Recognition.

Luc LUSSIER and Alain BIEM

2000.4.28

ATR人間情報通信研究所

〒619-0288　京都府相楽郡精華町光台2-2-2　TEL: 0774-95-1011

ATR Human Information Processing Research Laboratories

2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
Telephone: +81-774-95-1011
Fax　　　: +81-774-95-1008

Channel Distortion Equalization for Robust Speech Recognition

Luc Lussier and Alain Biem

April 25, 2000

Abstract

This report proposes an approach in speech recognition for handling mismatches in microphones that occurs between the training and the testing phase. Typically, it is assumed that the microphone used for training displays different characteristics from the microphone used in the testing phase. The proposed algorithm estimates a bias or distortion from the silence portion of the utterance. Assuming that a microphone acts as a filter on the incoming speech, the estimated distortion reflects the discrepancy between each microphones. Equalization is achieved by removing the estimated bias from the incoming speech.

1 Introduction

A fundamental problem in speech recognition that is yet to be solved is the degradation in performance of a speech recognizer when it operates under conditions that are different from the ones assumed in the training phase. These conditions include the environment in which the system operates and the transmission channel that carry the input speech to the recognition system. Both the environment and the transmissions channels affect the input speech in different ways. The environment may affect the input utterance in the form of an additive (background noise) and convolutional distortion (reverberation). Also, in an extremely noisy environment, an other phenomenon known as the Lombard effect, can further degrade the performance of the recognition process. As for the degradation of the input speech introduced by the transmission channel, it includes effects from the transducer (the microphone or the telephone handset) and the transmission lines. These effects are usually modeled as a filtering process being applied on the incoming speech signal. In this study we focus on attenuating the effects of the transmission channel on the performance of a speech recognizer.

Channel-robust speech recognition is of the utmost importance in state of the art speech recognizer design, because they are expected to operate with a wide range of transducers with varying transmission quality, including ordinary telephone, cellular telephone, speaker phone and a variety of microphones that possess different physical and acoustic characteristics. In addition, as previously stated, speech may be surrounded by a variety of environmental noise. From a practical viewpoint, such as price and convenience, users of a speech recognition system may simply not want to be restricted to use a particular microphone. For example, a system such as the HIP MECS system developed using a particular set of telephone handsets, may be expected to perform reasonably well when ported to an environment that uses a different set of microphones and telephone handsets.

Cepstral features are currently the most popular speech representation being used. The effect of the transducer differences between training and testing can be viewed as distortion in the cepstral domain. Various methods have been proposed to estimate and remove (or equalize) this distortion. This process usually takes the form of a bias removal or a filtering of the feature-vector time series. Cepstral Mean Subtraction (CMS) and RASTA processing are the most widely known representatives of these methods.

CMS computes the average cepstrum vector from the input utterance and then removes this average from the incoming speech cepstra. The assumption made here is that the mean of the cepstral vector is the effect of the distortion induced by the transmission channels. This is based on a theoretical result which states that the asymptotic means of a cepstrum is zero, for cepstrum indices higher than zero [1]. Thus, a non-zero cepstral mean is due to the effect of the transmission channels. CMS has been quite effective in a wide variety of tasks [3], including isolated word recognition as well as continuous speech recognition. However, CMS is not suited for real-time application and has to be applied to both the training and the testing data. The use of short-term cepstral averages has also been proposed, with limited results.

RASTA processing [2] acts as a band-pass filter applied to either the filter-bank log energies or the cepstral domain. It suppresses slow variations in the time trajectories, which are assumed to be due to channel effects and high characteristics caused by system artifacts.

Since both RASTA processing and CMS must be applied during both training and testing, they are not as useful when the task is to design a channel-robust system from already trained acoustic models that have not been previously processed. In this context, compensation techniques, such as SNR-Dependent Cepstral Normalization (SDCN) [4] or techniques that rely on the signal to noise ratio, or on other forms of environment models have been proposed, [3] , but at a high computational cost. Furthermore, these methods require a priori knowledge of the testing environment for estimating the SNR.

This report proposes a method that does not require a priori knowledge of the testing environment. It is assumed that a system, such the ATR HIP MECS system, has already been built, using a particular set of telephone handsets. The goal is to make the system robust to utterances that come from previously unknown transducers. The method adopted here is to use some part of the incoming speech signal that is assumed to be silence and then to estimate the distortion using the discrepancy between the silence contained in the input speech and the model.

2 Theoretical formulation

2.1 The channel distortion model

We adopt a simple model of the channel distortion. In the power spectral domain, the speech that comes from the transducer can be expressed as :

$$X(\omega) = H(\omega)S(\omega) + N(\omega)$$

where $X(\omega)$ represents the distorted and noisy speech spectrum, $S(\omega)$ is the power spectrum of the original speech, $H(\omega)$ is the transfer function of the transmission line, and $N(\omega)$ is the power spectrum of the additive noise. This relation can be expressed in the log-domain as

$$\log[X(\omega)] = \log\left[H(\omega)S(\omega) + N(\omega)\right]$$

which, by making use of the relation $\log(a+b) = \log(a) + \log(1 + \exp(\log(b) - \log(a)))$, can be expressed in the cepstral domain as,

$$c_x = c_h + c_s + r(x, h, s)$$

with

$$r(x,h,s) = \mathrm{DCT}^{-1}\left[\log\left(1 + \exp(c_x - c_s - c_h)\right)\right]$$

where c_x, c_s, and c_h are respectively the cepstral representation of the distorted speech, the clean speech, and the channel transfer function. DCT is the discrete cosine transform, which was previously applied to log energies to generate cepstral parameters. The term $r(x, h, s)$ depends on the speech and on the channel characteristics and can be assumed to have zero mean. Based on these assumptions, the mean of the speech that come from the transducer are

$$\mathrm{E}[c_x] = \mathrm{E}[c_s] + \mathrm{E}[c_h].$$

The previous equation has the following significance. When the model is created using a given transducer, represented here by c_h, the Gaussian means, within an HMM framework, are the sum of the mean of clean speech and the characteristics of the channel distortion. The same equation is valid, when testing from an unknown utterance x' that comes from a different microphone h' :

$$\mathrm{E}[c_{x'}] = \mathrm{E}[c_{s'}] + \mathrm{E}[c'_h].$$

For s representing a phonetic signal characterized by an HMM model and for s' having the same phonetic label as s, we can assume $\mathrm{E}[s] = \mathrm{E}[s']$. This means that the discrepancy in microphones equals the discrepancy between the cepstral means of a reference model and a testing signal, representing the same phonetic unit. That is,

$$\mathrm{E}[c_x] - \mathrm{E}[c_{x'}] = \mathrm{E}[c_h] - \mathrm{E}[c_{h'}].$$

The above equation lays the ground for a method of estimation of the channel distortion without actual knowledge of the channels characteristics. The difference in each channel is directly estimated from the incoming signal and the model, e.g., by estimating the difference $\mathrm{E}[c_x] - \mathrm{E}[c_{x'}]$. The difficulty in this method comes from the fact that, during testing, the label of the incoming speech is unknown before decoding is made by the system. One method to overcome this difficulty is to make use of a noise model and then assume that the first few frames of the incoming speech are mainly constituted of noise.

2.2 Algorithm description

The proposed algorithm is as follows. We assume a situation where a system has been built with previously stored HMM models. We suppose that there is a silence model available, estimated during the HMM training phase. The algorithm consists of three steps.

In the first step, the system builds a silence model at run time based on the first few frames of each utterances.

During the second step, the system computes the distortion c_d by comparing the silence model stored in the HMM model and the silence model estimated at run time. That is,

$$\mathrm{c_d} = \mathrm{E}[c_x] - \mathrm{E}[c_{x'}], \tag{1}$$

where as previously stated, $c_{x'}$ is the cepstral representation of the incoming signal x' during the actual use of the system.

In the last step, the incoming signal x' is transformed into a signal that bears the same statistical characteristics as the model. This is done by adding the previously computed distortion d to the incoming utterance x' :

$$c_y = c_{x'} + c_d.$$

The means $\mathrm{E}[c_y]$ of the cepstral representation of the transformed signal y is equal to the mean of the model of the same label, as shown below:

$$\begin{aligned} \mathrm{E}[c_y] &= \mathrm{E}[c_{x'}] + c_d \\ &= \mathrm{E}[c_{x'}] + \mathrm{E}[c_x] - \mathrm{E}[c_{x'}] \\ &= \mathrm{E}[c_x]. \end{aligned}$$

Consequently, by adding the bias d to the incoming speech signal, in the cepstral domain, the transducer channel difference between testing and training is attenuated.

Notice that all the above assumptions have been carried in the ideal situation, where the true means of the signal is available. In practice, we are far from knowing what the true mean value is. Furthermore, we have assumed that the term $r(x, h, s)$ is of zero mean, which may not be strictly true. The resulting distortion is thus a biased approximation to the true distortion. In practice, we must also smooth the unavoidable effect of this unbiased estimation. This is done by making use of a regularization term to account for the effect of the unbiased estimation. That is, we perform

$$c_y = c_{x'} + \alpha c_d, \tag{2}$$

where α is the regulating parameter.

(2) does not take into the account the variance of the models, but solely assumes that the distortion depends on the means and that the variances do not have any role to play. Again, accounting for variance can ease the unbiased estimation of the distortion. The method proposed here to account for the variance is to alternatively transform the feature by the following iteration:

$$c_y(i) = c_{x'}(i) + \alpha\sigma_i c_d(i), \tag{3}$$

where $c_y(i)$, $c_{x'}(i)$, $c_d(i)$are the i-th cepstral component of cepstrum vector c_y, $c_{x'}$ and c_d, respectively. σ_i is the i-th element of the covariance matrix, assumed to be diagonal.

3 General testing considerations

The channel distortion equalization algorithm implemented in the MECS system uses many parameters to determine its behavior. Because of the dimensionality of the testing space, some choices have been made to fix all of the other available parameters so that only the parameter α found in (2) and (3) would vary. Some of the other parameters are known to be sub-optimal in a given context, but trying to optimize all parameters all the time would be a computationally impossible task. Here are two relevant parameters for which the value was fixed experimentally:

Number of states from the HMM silence model used to represent the silence	3 out of 3
Number of frames taken at the beginning of the speech sound file that are considered to be silence	6

4 Test on ATR directory assistance task

The method previously described above, was applied to the ATR directory assistance task. The purpose of this task is to recognize Japanese names, within the ATR labs, uttered through a telephone handset and to forward the call to the corresponding persons if desired. In this particular application, the system was already trained and the challenging task is to make the system more robust to unknown microphone channels.

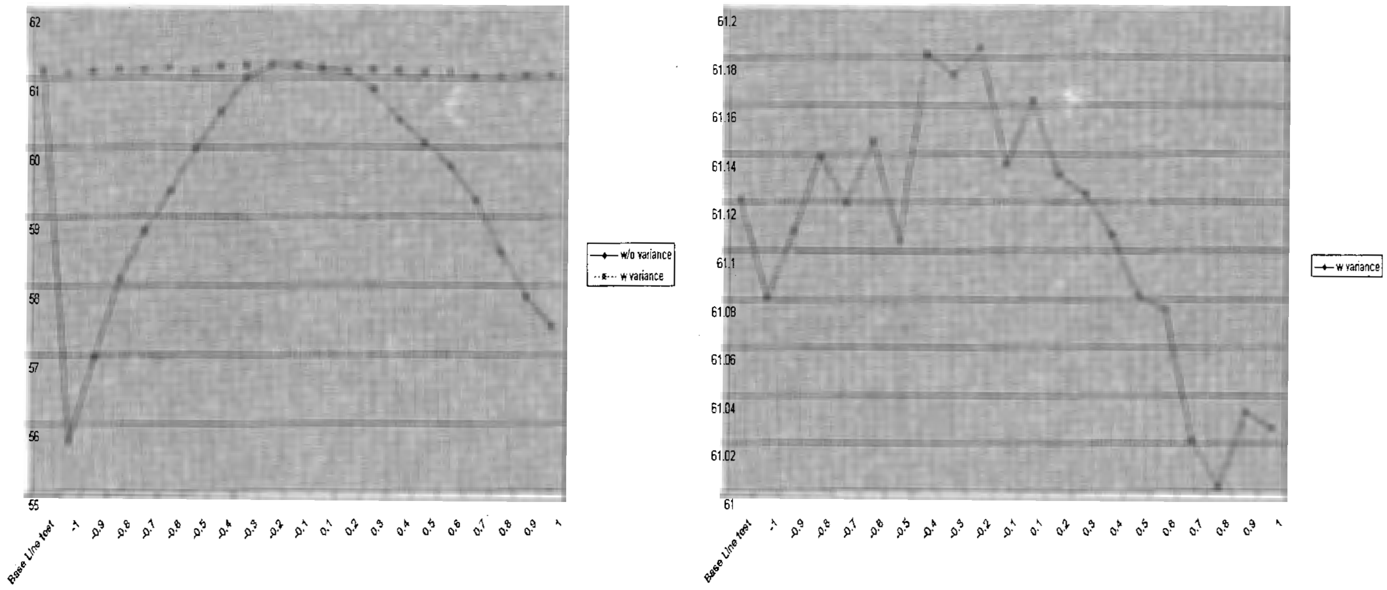

Figure 1: Effect of α. System trained and tested on ATR-Talk task. Left: without considering the variance. Right: considering the variance.

4.1 Experimental conditions

Based on the ATR directory assistance task, three different tasks were used to test the channel distortion equalization method. In all cases, the system was trained with the ATR Task and the test was performed on the ATR Task, MCD (high microphone quality) and MCD (low microphone quality) test tasks.

In each case, two tests are performed. The first test illustrates the effect on the results of the variation of the α parameter found in (2) while the second test shows the impact of the introduction of the variance as in (3).

4.2 Results

As shown on the left side of Fig. 1, when the algorithm is applied to a testing task which is identical to the training task, the best results are obtained for an α parameter close to 0. This result indicates that even when both physical channels are identical, the algorithm estimates a distortion d which is large enough to deteriorate the recognition. This is why the results are the best when the α parameter is close to 0 as this is equivalent to not performing any equalization. The right side of Fig. 1 illustrates the effect of considering the variance on the recognition. In this case, considering the variance visibly reduces the effect of the α parameter. The results in this case vary less than 0.2% for the whole range of testing.

In the case of Fig. 2 where the testing task represent a channel of inferior quality than the training task, the best results are achieved for $\alpha = -0.9$ and are degraded on the other side of the testing range. When the test is performed using (3) the curve display an identical tendency but an inferior result is obtained in the best case. In this test, the best results present improvement of almost 50% over the base line test. However, this improvement does not translate directly into a greater number of correctly recognized phonemes as MECS also takes into consideration the number of phoneme mismatches, deletions and insertions in its computations.

The last ATR directory assistance task presents a test task for which the channel is of higher quality than for the training task. The results in Fig. 3 are different than what was first expected since the best results are obtained for $\alpha = -0.7$ when a positive value of α was expected to yield the best result. Also, for this test, the result of (3) is not as clear as in the two previous experiments. A peak at $\alpha = 0.4$ gives a recognition value which is still higher than the base line result.

4.3 Observations

As shown on the left side of Fig. 1, there seems to be a problem with the estimation of the distortion in (1). Of course, since we arbitrarily label the first 6 frames from each speech file as being silence, we can expect that this probably introduces frames that have a non silence content into the algorithm. The use of the variance also produces interesting results as in the right side of Fig. 1 where it almost makes the system invariant to the different values of α. On the whole test space, the percentage of recognition varies less than 0.2% from the base reference value when the variance is taken in consideration while the recognition degrades more than 5% in the worst case when the variance is not considered.

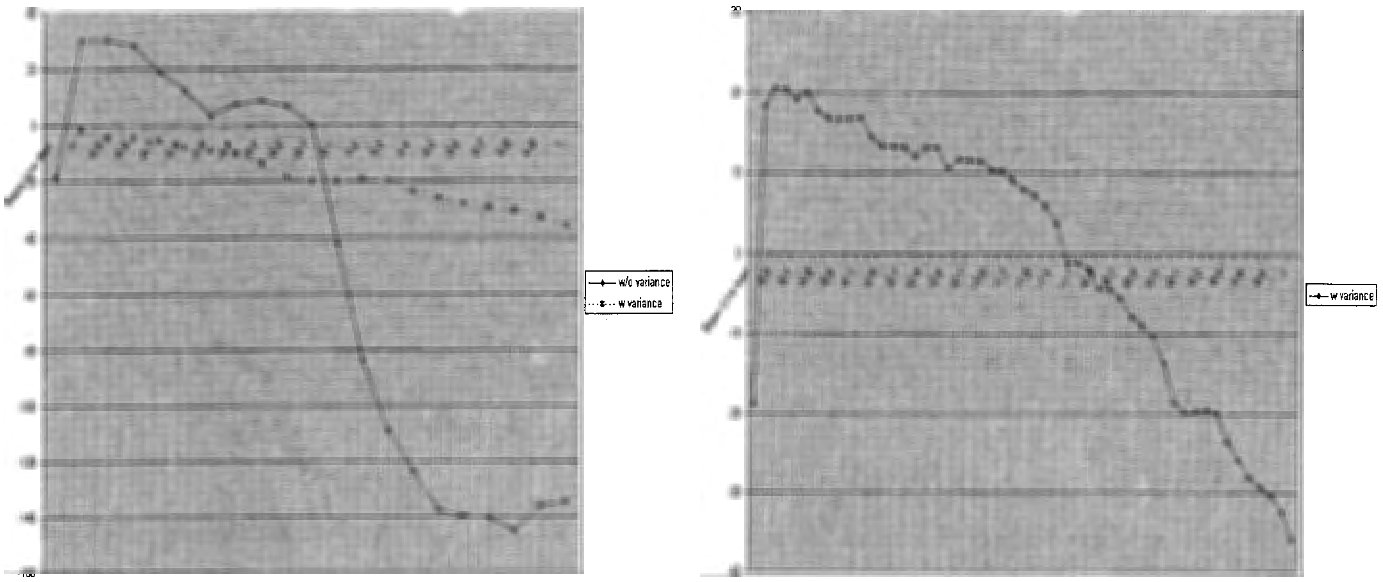

Figure 2: Effect of α. System trained on ATR-Talk task and tested on MCD (low microphone quality) task. Left: without considering the variance. Right: considering the variance.

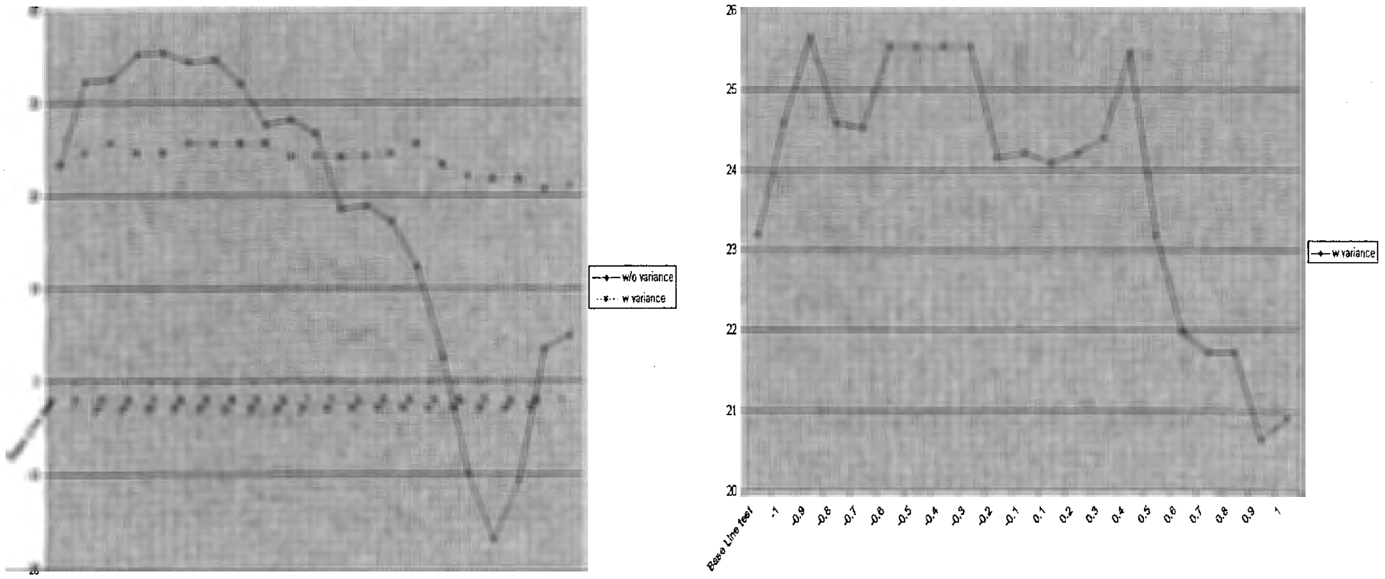

Figure 3: Effect of α. System trained on ATR-Talk task and tested on MCD (high microphone quality) task. Left: without considering the variance. Right: considering the variance.

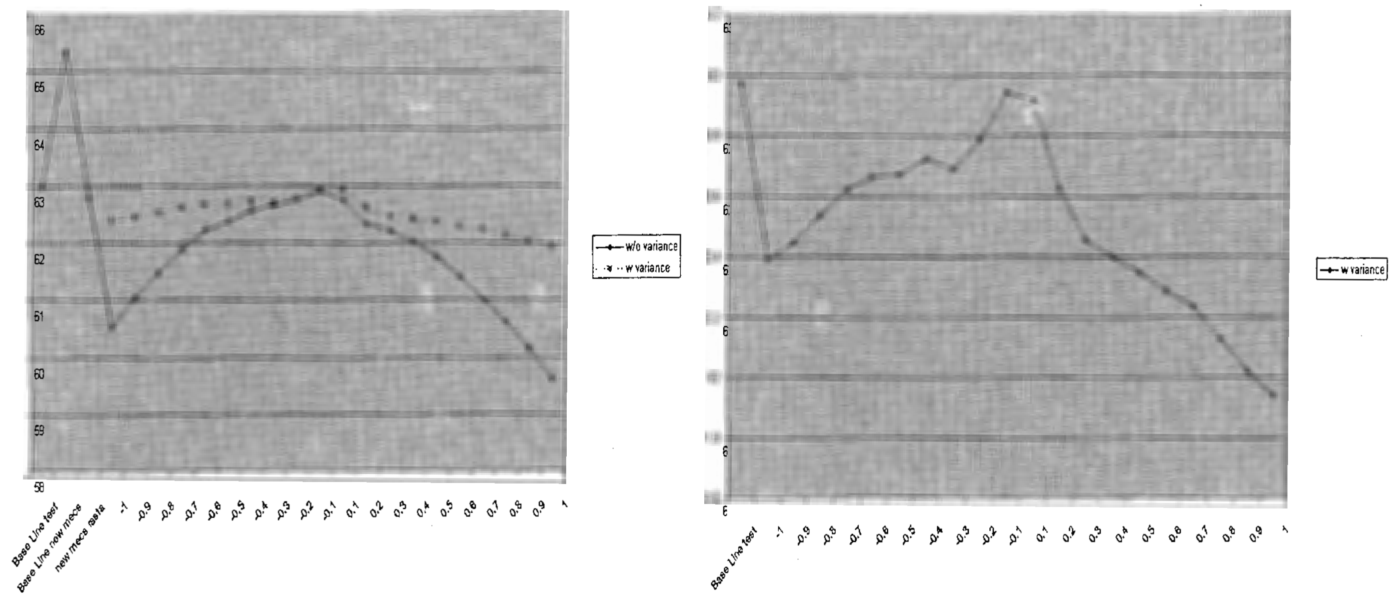

Figure 4: Effect of α. System trained and tested on the TIMIT task. Left: without considering the variance. Right: considering the variance.

5 Test on TIMIT and NTIMIT task

Experiments were also made with the TIMIT and NTIMIT task. The TIMIT database consists of relatively clean speech recorded in a noise-free environment with the purpose of providing a framework for continuous speech phoneme recognition. NTIMIT was obtained by transmitting the TIMIT utterances through a telephone line. The strategy used here is to train the model using one database and then perform testing using the other database. As the NTIMIT database is deliberately a channel distorted version of the TIMIT database, the two databases provide a good framework for testing the accuracy of the proposed approach.

5.1 Experimental conditions

Based on the TIMIT and NTIMIT tasks, two different experiments were performed.

In the first experiment, for each combination of testing and training tasks, two tests are performed. The first test illustrates the effect of the variation of the α parameter found in (2) while the second test shows the impact of the introduction of the variance found in (3).

In the second experiment, a test was performed for each combination of testing and training tasks to better understand the effect of the algorithm on recognition. During these tests, the algorithm applies corrections only to specific elements of the feature vector used in the recognition system.

5.2 Results

5.2.1 Experiment 1

As expected, Fig. 4 shows that when the training and the testing tasks are similar, the best results are obtained when α is close to 0.

In the following test, Fig. 5, the best results are obtained for values of α equal to 1.6 and 3.4. Both of these values are outside of the initially expected possible range for the α parameter, which was originally thought to be included in the $[-1, 1]$ interval.

Again, as shown in Fig. 6, the best results are obtained for a value of α close to 0.

Fig. 7 presents a unique behavior as the best results on the left side of the figure are obtained for a positive value of α and on the right side of the figure for a negative value of α. This is the only case where the use of the variance significantly modifies the location of the maximum. Again, it should be observed that when the variance is considered, the results on the whole test range vary less than 1%.

In Fig. 4 to Fig. 7, the left side graphics second and third value from the left, which are labelled "Base Line new mecs" and "new mecs rasta", illustrate the results produced by the latest version of MECS. This recent version of MECS has

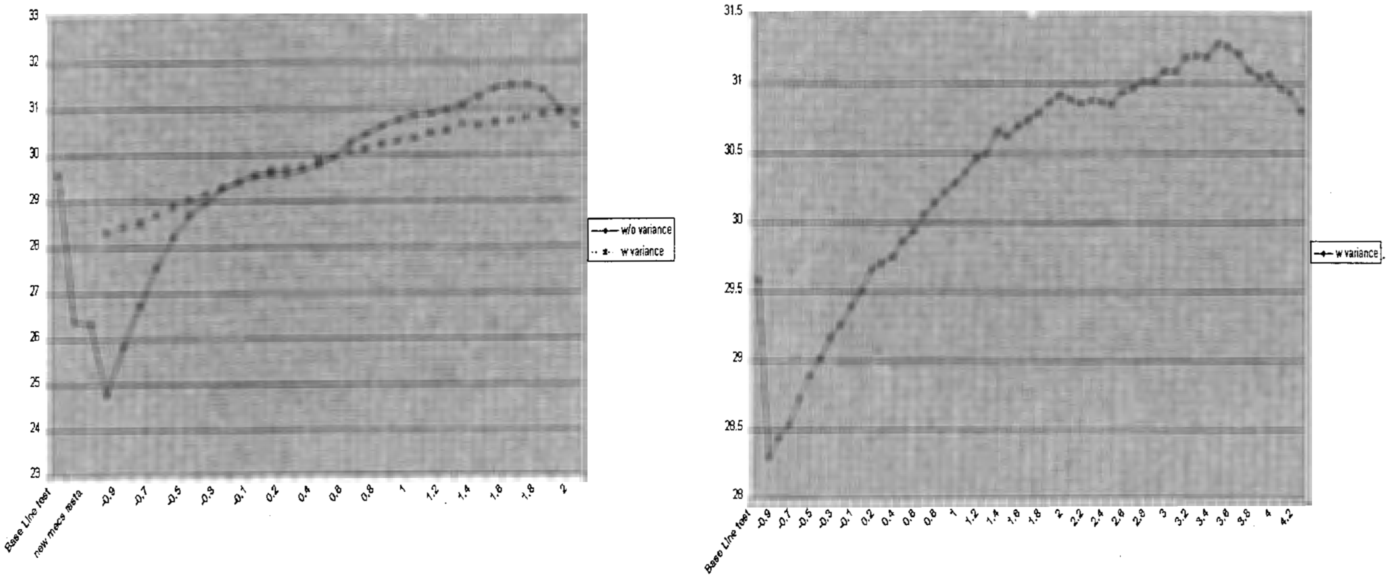

Figure 5: Effect of α. System trained on the TIMIT task and tested on the NTIMIT task. Left: without considering the variance. Right: considering the variance.

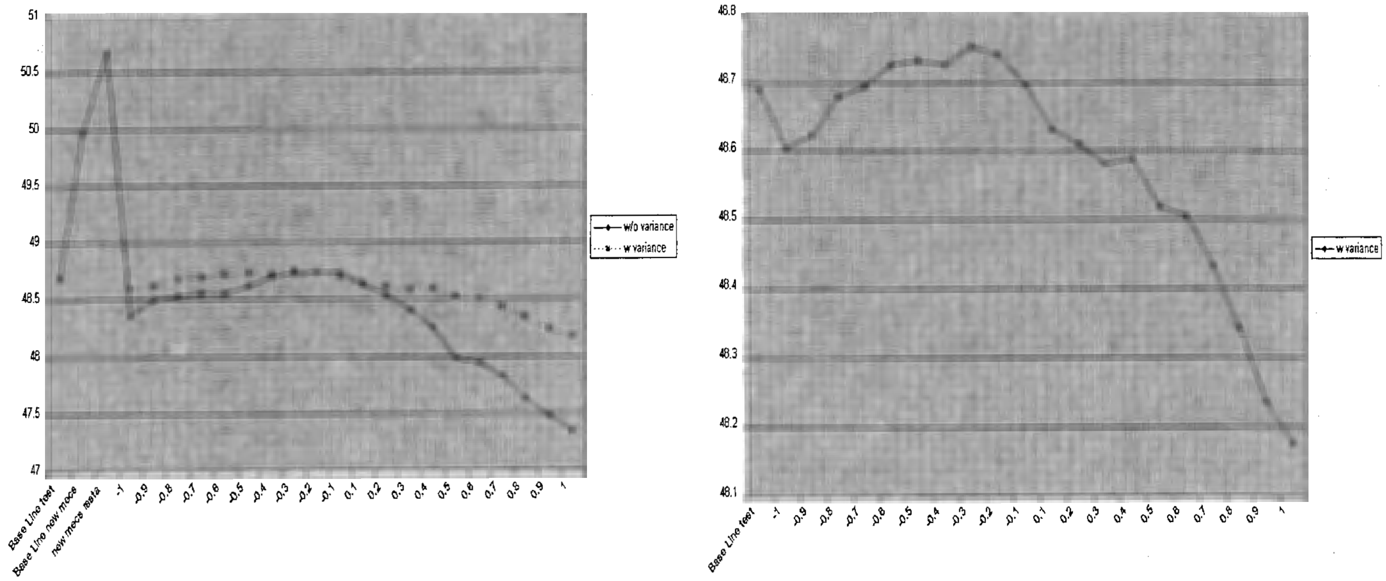

Figure 6: Effect of α. System trained and tested on the NTIMIT task. Left: without considering the variance. Right: considering the variance.

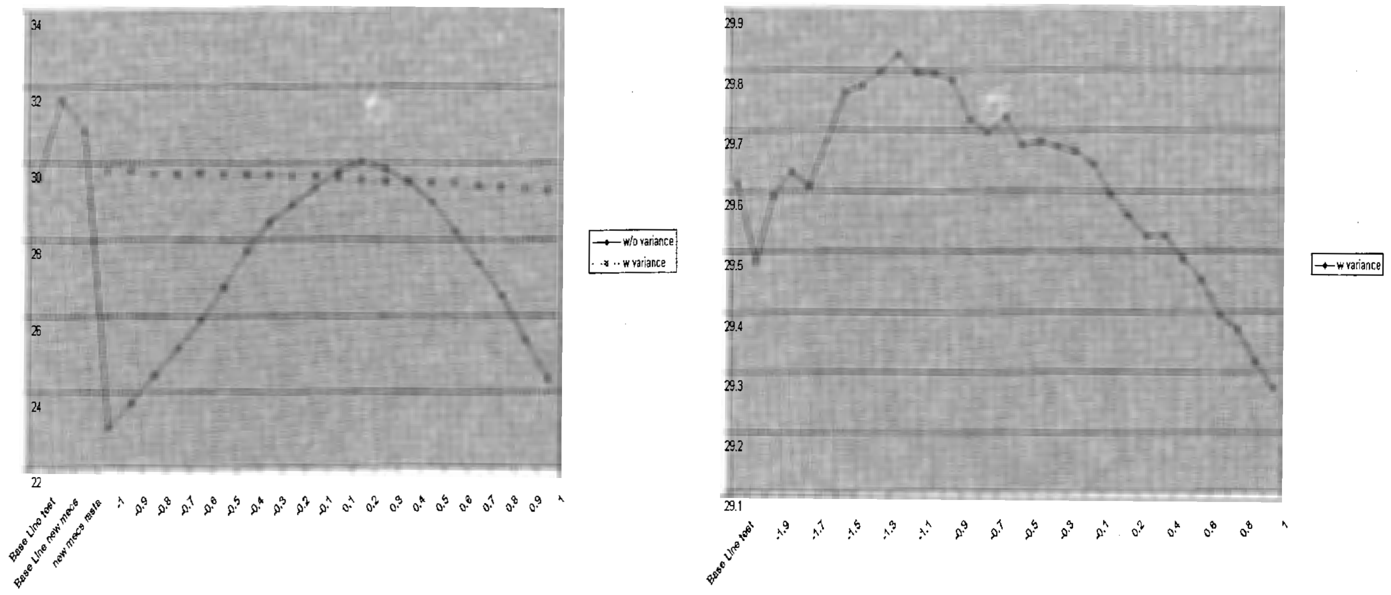

Figure 7: Effect of α. System trained on the NTIMIT task and tested on the TIMIT task. Left: without considering the variance. Right: considering the variance.

undergone a few internal changes from the version of MECS used to perform all the channel equalisation tests but since the older version could not perform RASTA processing, the newer version had to be used. For comparison purpose, a base line test was also produced with the newer version of MECS.

The comparison of the results obtained between the base line and the RASTA tests show that only in the case of a system trained and tested with the NTIMIT task, Fig. 6, does the RASTA processing enhance the recognition results while in the other cases, the RASTA processing result in a reduction of the recognition percentage.

5.2.2 Experiment 2

Of all the tests performed, the following tests show the most surprising results. By applying the channel equalization only to specific elements of the vector, the goal was to observe the effect of the equalization on each element. Again, all possibilities could not computationally be covered because of the size of the testing space, which has 39 dimensions. The energy, δ-energy and $\delta\delta$-energy each represent one dimension while the MFCC, δ-MFCC and $\delta\delta$-MFCC each has 12 dimension. Also, energy and MFCC were evaluated separately and for the MFCC, each 12 dimensions vector section was considered as a single element to form seven categories in each case. As for the α parameter, it was arbitrarily set to -1 or 1, a choice known to give sub-optimal results.

Both tests for the TIMIT trained system, Fig. 8, show corresponding results and show that the equalization process has very little impact when applied to the cepstral region of the vector while its application to the energy region shows a more important effect.

On the other hand, when testing with the NTIMIT trained system, Fig. 9, both the energy and the cepstral components of the vector influence the results visibly, considering a scale of 1% and 0.1% for those figures.

5.3 Observations

Once more, we could observe in the experiment 1 that considering the variance has the effect of minimizing the impact of varying α. As for experiment 2, it shows that the actual effect of the algorithm is quite different from its intended purpose. Even if the algorithm has some impact when applied to the cepstral components of the vector, its effect on the energy seems more important, specially in the case where the system was trained with the TIMIT task.

6 Conclusion

In this report, an approach for handling the mismatch in the microphones used during the training and the testing phase was proposed. As a first step in this direction, a simple model of the channel distortion was used. Based on this model and assuming ideal conditions, the silence representation found in a trained HMM model used by the ATR MECS system was

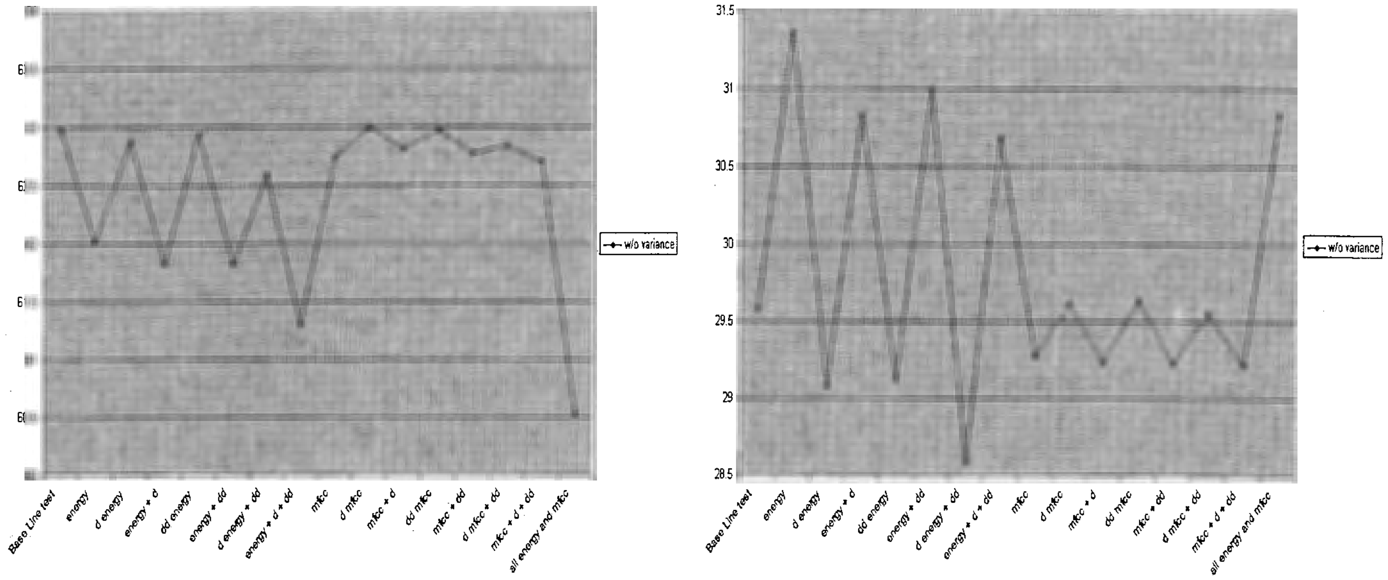

Figure 8: Application of the normalization to specific elements of the feature vector on a system trained on the TIMIT task. Left: TIMIT task is used for testing ($\alpha = -1$). Right: NTIMIT task is used for testing ($\alpha = 1$).

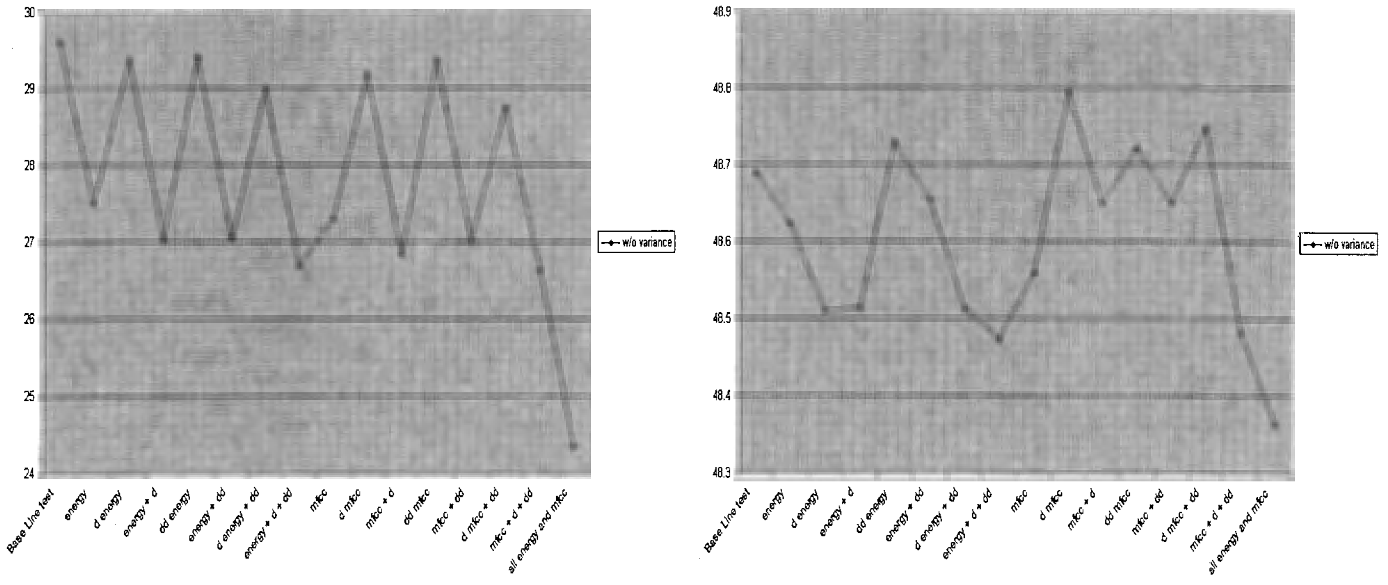

Figure 9: Application of the normalization to specific elements of the feature vector on a system trained on the NTIMIT task. Left: TIMIT task is used for testing ($\alpha = 1$). Right: NTIMIT task is used for testing ($\alpha = -1$).

compared at runtime with what was expected to be frames of silence taken from input speech. The distortion measured between the model and the input speech was then used to correct the rest of the speech utterance with the results presented previously.

All the tests performed show that when the training and testing tasks are similar the algorithm can not improve the recognition. Unfortunately, our measurement of the distortion also seems to introduce a lot of error.

The second experiment performed on the TIMIT and NTIMIT test set showed that while we first thought that this algorithm would especially affect the cepstral component of the vector, the energy components were also significantly influenced. This might indicate that an approach like this one could be used by the system to normalize the incoming speech according to various components found in the model. In this way, the system might be more aware of certain constant differences found between the input speech and the model used for the recognition.

Also, the scaling parameter, α could be further optimized automatically by the system by first trying to recognize an unknown phoneme without any equalization and then trying to identify the same phoneme by using various values of α and observing the variation of the degree of confidence of the recognition. This could lead to a system that would use different values of α for recognizing different phonemes. However, such an approach would lead to a higher demand of computation power.

An other element which is crucial to the performance of this system is to chose which frame is to be used by the system as a frame containing silence. In the presented tests, the six first frames of each speech utterance were arbitrarily used. This method presents two significant problems. First, the population of samples acquired in this way is relatively small and especially during the recognition of the first few utterances, the quality of the considered frame can have a significant impact. Second, the actual content of the silence labeled frames was not rigorously controlled and could have contained undesirable noise or speech. If the speech recognition system could identify which frame to consider as silence by looking at the content of each frame instead of blindly using some specified frames, the results might be more reliable. This would also allow this algorithm to be used more easily in a real-time situation where the microphone is always transmitting input to the system.

References

[1] S. Furui. Cepstral Analysis Techniques for Automatic Speaker Verification. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 29:254–272, Apr. 1981.

[2] H. Hermansky. Rasta-plp speech analysis technique. In *ICASSP 92*, pages I.121–I.124, 1992.

[3] J. C. Junqua and J. P Haton. *Robustness in Automatic Speech Recognition.* Kluwer Academic Press, 1996.

[4] F.-H. Liu, R. M. Stern, X. Huang, and A. Acero. Efficient cepstral normalization for robust speech recognition. In *Proc. ARPA Human Language Technology Workshop '93*, pages 69–74, Princeton, NJ, March 1994. distributed as *Human Language Technology* by San Mateo, CA: Morgan Kaufmann Publishers.