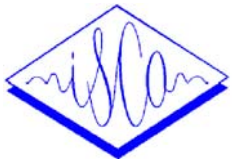


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ON THE USE OF TIME ALIGNMENTS FOR NOISY SPEECH RECOGNITION

Y. Hauptman
Natural Speech Communication
33 Lazarov st., Rishon Le'zion
Israel
galsafe1@internet-zahav.net

Y. Bistritz
Dept. of Electrical Engineering
Tel Aviv University
Israel
bistritz@eng.tau.ac.il

ABSTRACT

One of the basic aspects of modern pattern matching algorithms used in speech recognition is time-alignment. The use of time-alignment is essential for offsetting speaking rate variations, which is an inherent property of the speech signal. It is known that time-alignment contributes to increased accuracy in speech recognition. However, a key question is whether time-alignment information still contributes to recognition accuracy in highly degraded speech. In this paper we examine the robustness of time-alignment information by introducing a robustness indicator. Isolated words recognition experiments with and without time alignment (using DTW and VQ respectively) are used and to illustrate the issue.

1. INTRODUCTION

Speech recognition applications require accurate recognition over noisy environments such as telephone and cellular lines. In these environments, it is necessary for the system to maintain accurate recognition for highly degraded speech. Methods and algorithms to increase the system's robustness to noise include model adaptation, the use of robust speech features, and speech filtering tools. However, in severely noisy conditions, some parts of the speech might be completely corrupted, regardless of the robustness method used. In such cases, the recognition accuracy drops when these corrupted portions are used because the local similarity scores in the pattern-matching algorithm become invalid. In fact, it has been suggested that irrespective of the recognition mechanism, the overall accuracy will be higher when these corrupted portions are not taken into account by the recognizer [1],[2]. Currently, we present results showing that the adverse influence of the corrupted portions is not localized to just the ruined frames but affects the whole time-alignment path.

In section 2, two well-known recognition algorithms are briefly presented; one that uses time-alignment, and one that does not. In section 3, we compare these recognition algorithms by introducing a new robustness indicator. In section 4, the performance of these two algorithms is compared for an isolated word recognition (IWR) task in additive noise environment.

2. SPEECH RECOGNITION ALGORITHMS

Recognition algorithms are essentially means for calculating the similarity distance d_k between incoming utterance X and each one of the K reference utterances R^K:

$$d_k = D(X, R^k) \quad ,k = 1,..., K \qquad (1)$$

where X, consists of a sequence of N feature vectors : $X = \{X_1, X_2, ..., X_N\}$, and each reference utterance R^k consists of M_k feature vectors $R^k = \{R_1^k, R_2^k, ..., R_{M_k}^k\}$. The recognized utterance is the closest reference utterance R^r to the incoming utterance:

$$r = \underset{k}{ArgMin}\ d_k \qquad (2)$$

The similarity distance, D, is calculated by summing up local similarity distances between of sequence pairs of feature vectors in the two utterances by:

$$d_k = D(X, R^k) = \frac{1}{W}\sum d(X_i, R_j^k) \qquad (3)$$

where W is a length normalization factor and i,j are frames indices.

Recognition algorithms can be classified by the local similarity distance used and by the method of assigning frames from the two utterances. In the next section we describe two simple, well known, algorithms which are successfully used for isolated words recognition tasks.

2.1 The DTW Algorithm

The DTW algorithm [3] calculates the time alignment P_k between a test utterance and a reference utterance k. It is represented by a path, i.e. a sequence of pairs of integers, in the two dimensional grid:

$$P_k = \{(i_l, j_l), l = 1,..., L\} \qquad (4)$$

P_k is the path which minimizes the overall similarity distance between the utterances:

$$P_k = ArgMin \sum d(X_i, R_j^k) \quad (5)$$

Practically, the choice of the time alignment path, P_k, should follow some restrictions based on the physical nature of the speech signal itself. Such restrictions are

- End points region restrictions
- Monotonicity\orientation
- Local continuity
- Min\Max Slope limitation

Eq. (5) implies that every local distance (and therefore every frame) contributes to the path calculation and hence on the overall form of the path. This fact, together with the above path restrictions, imply that a local distance in a frame or several adjacent frames may actually shift the whole path location.

2.2 The VQ Algorithm

When using Vector Quantization (VQ) for isolated word-recognition [4], Eq.(3) becomes :

$$d_k = \frac{1}{W} \sum_i d(X_i, Q^k(X_i)) \quad (6)$$

where $Q^k(X_i)$ is the quantization operation defined by:

$$Q^k(X_i) = R_j^k \; ; j = ArgMin \; d(X_i, R_n^k) \quad (7)$$

Eqs. (6) and (7) are not restricted by time alignment considerations as is the case in DTW. Each frame affects the global distance and the recognition accuracy only through its local distance contribution.

3. ROBUSTNESS OF TIME-ALIGNMENT TO ADDITIVE NOISE

The robustness of time-alignment was investigated by comparing the relative recognition performance using DTW and VQ in various levels of additive noise. The two criteria that were used to measure performance were: 1) the robustness indicator and 2) recognition results. We chose DTW and VQ to represent recognition with and without time-alignment respectively.

3.1 Robustness Indicator

The recognition process in Eqs. (1) and (2), assumes that the similarity distance between a tested utterance and its correct reference template is smaller than its distance to all other reference templates. A recognizer is robust to noise if it can retain this relation also for degraded speech. It therefore makes sense to measure the robustness to noise of a recognition algorithm by the relative increase in the distance to the correct template that occurs when noise is added to the tested utterances.

Consider two versions of the same utterance: a 'clean' (high SNR) version X , and its degraded version $\tilde{X}$. We define the sameness acceptance measure (SAM), at a given SNR level, as the average ratio, over all the tested utterances, of the similarity distances between the degraded version of the utterance from it's appropriate reference template, and the similarity distance of a 'clean' version of the utterances to its appropriate reference template:

$$SAM = \underset{X}{average}\left\{\frac{D(\tilde{X}, R^k)}{D(X, R^k)}\right\} \quad (8)$$

where,

$$k = ArgMin \; D(X, R^j) \quad (9)$$

This measure may be used to compare the performance of different pattern-matching algorithms for noisy speech. The SAM is greater than 1 for noisy speech and increases as the level of noise increases. Lower SAM values indicate a more robust algorithm.

3.2 Experimental Setting

The SAM indicator was used to compare the relative robustness of speaker dependant IWR by VQ and DTW. We conducted experiments using the following data base and experimental setup:

- Database: We used the TI-46 digits database that consists of 16 speakers (8 – male, 8 – female). For each speaker 10 repetitions were used for training and 16 for testing.
- Speech processing: Cepstral coefficients derived from frames of 20msec with 50% overlap of signals dawnsampled to 8Khz
- Training: The reference utterances for DTW were obtained by averaging all the repetition of utterances for each digit and for each speaker. The training of the VQ used K-means to assign a codebook of size 64 for each digit and each speaker.
- Degradation conditions: In order to tune the degradation level we used additive white noise. The noise sequence was amplified by a constant before added to the clean speech signal to obtain the desired SNR value.

3.3 Robustness Indicator – Experimental Results

In the following are some of our measurements of SAM for VQ and DTW based IWR for different values of SNR using the above experimental setting and the following additional conditions. LPC of order 16 was obtained for each frame and used to derive 18 LPC-cepstrum coefficients as feature vectors for both VQ and DTW. (Other orders were too examined but will not be reported here.) The results of measuring the SAM are shown in Figure 1. It is seen that VQ tends to maintain higher robustness at SNR range of 20dB to –3 dB. This means that the similarity distance between the reference templates and noisy incoming utterances degrades less in this range without time-alignment.

Therefore recognition may be expected to perform better without time alignment in this wide range of SNR's. We shall later address the reason why at very low SNR's (below –3dB) DTW seems to be more robust.

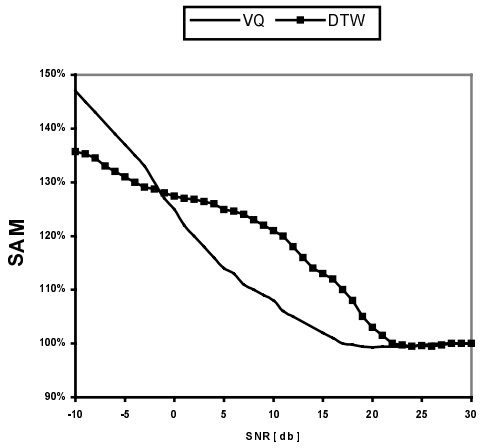


Figure 1. Sameness acceptance measure (SAM) results for various noise levels.

4. RECOGNITION EXPERIMENTS

For a different perspective on the effect of time alignment on recognition we compared the recognition accuracy of DTW and VQ using the same experimental setting as in the SAM measurements in §3.2. The tests were carried out using several cepstral feature vectors based on linear prediction (LPC-CEP), perceptual linear prediction (PLP-CEP) and Mel-cepstrum (MFCC), (cf. e.g. [8]). For each feature type, the feature vector size was chosen to optimize the performance at 0 dB. The results of these experiments are shown in Figures 2,3 and 4. These figures show that independently of the chosen feature vector, there exists a wide mid range of SNR's for which the omission of time-alignment (using VQ) improves performance. These results are consistent with the previous SAM measurements.

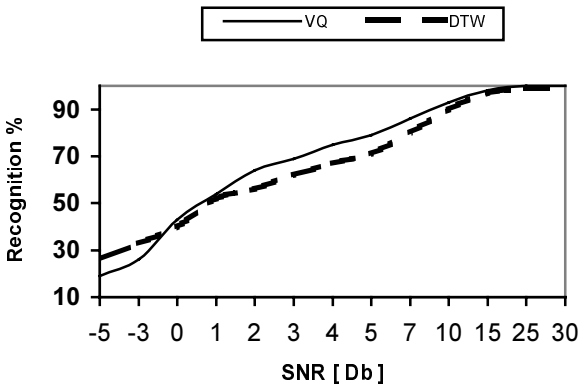


Figure 2. Simulation results using DTW and VQ using LPC-CEP (order 20)

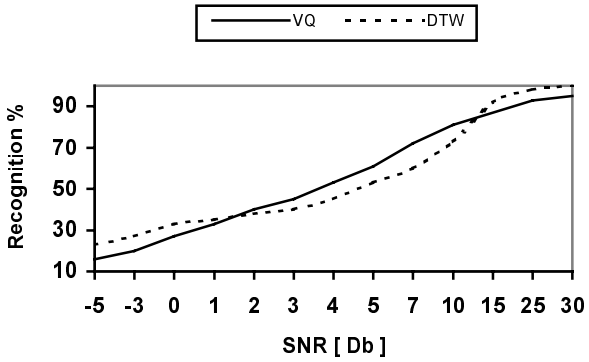


Figure 3. Simulation results using DTW and VQ using PLP-CEP (order 12)

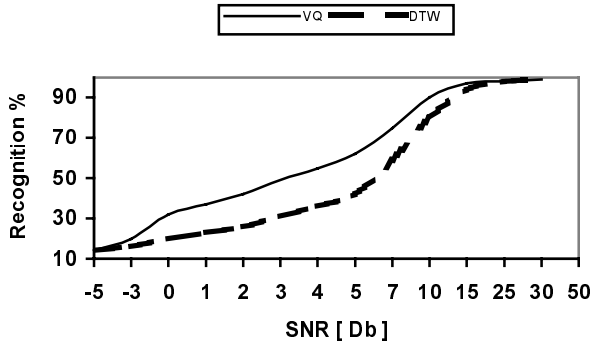


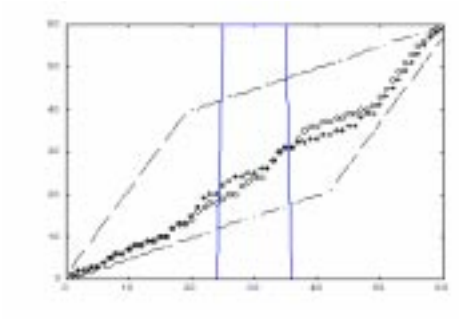
Figure 4. Simulation results using DTW and VQ using MFCC-CEP (order 22 with 13 filters)

5. DISCUSSION

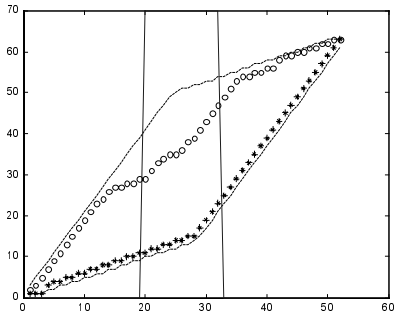
This study shows that recognition without time-alignment improves performance of IWR in an additive noise environment for SNR levels in the mid-range (~ 0 to 15 dB). For high SNR values, as expected, time alignment is beneficial for recognition accuracy. Surprisingly, at very low SNR values, DTW again wins with better accuracy. This may be explained by the fact that at very high noise levels most of the utterance is so corrupted that the only left reliable information was some timing information (e.g. the utterances lengths that were pre set in this IWR experiment) that DTW, unlike VQ can capture.

The time-alignment path is calculated to minimize the overall similarity distance between the two utterances. Therefore, when local similarity distances fail, this can perturb the whole time-alignment path, decreasing the ability of less corrupted parts of the utterance to contribute to the recognition. Figure 5 illustrates this phenomenon. Figure 5(a) describes alignment of a tested 'EIGHT' to a template 'EIGHT'. The path marked with '*' corresponds to a clean tested utterance. We then added white noise at 0 dB SNR to only a few frames in the middle of the utterance (between the two

vertical mark lines). The resulting aligned path is marked by 'o'. Figure 5(b) repeats the experiment with 'ONE' as the tested and 'FIVE' as the template utterances. It is seen that although only the middle part of the utterance is corrupted, the whole path shifts from its original location and as a consequence the contribution of the clean portions of the utterance will be diminished.



(a)



(b)

Figure 5. Effect of a corrupted segment of speech on the time-alignment path.

6. CONCLUSIONS

In this work the effect of additive noise on the time-alignment path was examined. The performances of a basic DTW algorithm and a VQ algorithm were compared. We demonstrated that the use of time-alignment, such as DTW, results in increasing the sensitivity of the recognition system to additive noise. Although some modifications on the basic DTW algorithm might increase robustness (e.g. [5],[6],[7]), they can not eliminate the inherent disadvantage in time-alignment. The better performance that VQ achieves for recognition of isolated digits in a wide range of noised speech compared to DTW was shown to be caused by the sensitivity of the time-alignment mechanism to noise. It was also demonstrated that the noise impacts the alignment beyond its local occurrence.

Using VQ alone, as in our IWR experiments, is not a viable solution for more general speech recognition tasks that today use mostly Hidden Markov Models with left-to right topology. However this topology functions like time alignment in DTW. This study suggests that some relaxation of the time-alignment constraints when the speech is highly degraded can improve the performance of a speech recognizer.

REFERENCES

[1] S. Ahmed and V. Tresp, "Some solutions to the Missing Feature Problem in Vision", in *Advances in Neural Information Processing Systems*, Volume 5, pp. 393-400 (S. J Hanson, J. D Cowan, and C. L. Giles Eds.), Morgan Kaufmann, San Mateo, 1993 .

[2] R. P. Lippmann and B. A. Carlson "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise", Eurospeech-97, pp. KN-37-40

[3] H. Sakoe and S. Chiba " Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. Acoust. Speech, Sig. Proc*., Vol. Assp-26, pp. 43-49. Feb 1978

[4] J. E. Shore and D. K. Burton, "Discrete utterance speech recognition without time alignment", *IEEE Trans. Information Theory*, Vol. IT-29, pp. 473-491, 1983.

[5] C. Myers, L. Rabiner and A. E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition", *IEEE trans. on Acoustics Speech and Signal Processing*, Vol. ASSP-28, pp. 623-635 , Dec. 1980.

[6] L. Rabiner , A. E. Rosenberg and S. E. Levinson Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition", *IEEE trans. on Acoustics Speech and Signal Processing* , Vol. ASSP-26, pp. 575-562 , Dec 1978.

[7] I. Shalom and A. Cohen, "On time warping algorithms for speech recognition ", Proc. 14th. Conv. of Elect. & Electron. Eng. , Tel-Aviv 1985

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