Time-Normalization Techniques for Speaker-Independent Isolated Word Recognition

S. Uma, V. Sridhar and G. Krishna
Department of Computer Science and Automation
Indian Institute of Science
Bangalore 560 012, India

Abstract

In this paper, we investigate various time-normalization techniques that are useful in the context of speaker-independent isolated word recognition. At the lowest level, we make use of LPC coefficients as the features to be nomalized. We discuss the various methods by which we can normalize these features. To begin with, we arrive at a typical number of frames associated with a word. Then, we normalize all the training and test data to this number of frames. Initial results bring out the point that normalization techniques help in reducing the number of patterns with which the unknown has to be compared.

1 Introduction

Normalization of speech parameters in some form or the other is essential for achieving speaker-independent isolated word recognition [1]. Number of training data necessary to achieve the required degree of accuracy may largely vary from normalization to normalization. For instance, the approach for speaker-independent isolated word recognition based on mean square polynomial classifiers described in [2] makes use of a large training data set. Even though in speaker-independent isolated word recognition systems training is performed "off-line", it is better to restrict the number of training samples to a manageable number.

In this paper, we propose approaches for time-normalization of speech parameters. The approaches have potential for restricting the comparison of the unknown patterns to a subset of available reference patterns. We also compare the proposed approaches, the approach based on dynamic time warping technique and the approach based on matrix norm.

2 Approaches for time-normalization

Time-normalization is a process of normalizing the duration of speech to a known duration of time. Such time normalization is meaningful especially in the context of isolated words as isolated words are always embedded by beginning and ending silences which can be detected and removed before the normalization. Normally, time-normalization of a pattern (a sequence of frames) is with respect to duration of another pattern [3]. Time-normalization can be either linear or non-linear. In linear time-normalization, the diagonal line of the rectangle formed by the two patterns corresponds to linear time alignment between the patterns. The well-known approach for non-linear time normalization of a pattern with respect to another pattern is DTW [4] and modifications to this approach have been suggested in the context of isolated word recognition in [5,6]. It is to be observed that the time-normalization of a pattern as described above is with respect to another pattern. Additionally, the patterns can be time-compressed prior to comparison. The stationarity of speech segments [7] is exploited by these techniques [8,9,10].

A variation in the time-normalization can be to time-normalize a pattern with respect to a standard duration instead of with respect to another pattern. In other words, normalization can be with respect to a standard duration of time for a given word that is in some sense independent of speakers. For instance, one can make use of vocal-tract models that view the vocal-tract as an open-ended tube with varying dimensions to account for the commonalties at the structure/process level rather than at the output/performance level. On the other hand, one can use statistical methods to arrive at the standard duration for a word. Normalization as described in the following helps not only to normalize two patterns to be compared but also to account for speaker-normalization.

A pattern is defined to be a sequence of frames where each frame is a (single-dimensional) vector. Let $\mathcal{C}$ be a collection of "reference" patterns. In the following, we describe different ways of defining a collection of reference patterns.

$\mathcal{C}$ is defined as follows:

$\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ where $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are k classes corresponding to k isolated words in the vocabulary.

$\mathcal{C}_i$ is defined as follows:

$\mathcal{C}_i = \{\mathcal{P}_{i1}, \ldots, \mathcal{P}_{im_i}\}$ where $\mathcal{P}_{ij}$ is the j^{th} pattern in the i^{th} class and there are m_i reference patterns in the class $\mathcal{C}_i$.

From the definition of the pattern, we have

$\mathcal{P}_{ij} = \{\mathcal{F}_{ij1}, \ldots, \mathcal{F}_{ijn_{ij}}\}$ and there are n_{ij} frames in the pattern $\mathcal{P}_{ij}$.

We define time-normalization as a function $\mathcal{N}$ from a set of patterns into a set of patterns. In particular, we are interested in the following two $\mathcal{N}$-functions.

Let $\mathcal{S}$ be a set of patterns.

0-8186-2920-7/92 $3.00 © 1992 IEEE

1. $\mathcal{N}^- : \mathcal{S} \rightarrow \mathcal{S}$ is defined as follows: $\mathcal{N}^-(\mathcal{P}) = \mathcal{P}'$ if $\mid \mathcal{P}' \mid = \mid \mathcal{P} \mid - 1$.

Note that $\mathcal{P}$ is a sequence of frames and $\mid \mathcal{P} \mid$ denotes the length of that sequence.

We define $\mathcal{N}^-$ compositions r times as follows:
$(\mathcal{N}^-)^r(\mathcal{P}) = \mathcal{P}'$ if $\mid \mathcal{P}' \mid = \mid \mathcal{P} \mid - r$
When $r = 0$, we have $(\mathcal{N}^-)^0(\mathcal{P}) = \mathcal{P}$.

In other words, there is a sequence of patterns $\mathcal{P} = \mathcal{P}_1, \ldots, \mathcal{P}_r = \mathcal{P}'$ such that
$\mid \mathcal{P}_{i+1} \mid = \mid \mathcal{P}_i \mid - 1$ for $1 \leq i \leq r-1$.

Definition of $\mathcal{N}^-(\mathcal{P})$:
Let $\mathcal{P} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_x\}$ be a pattern.
Let $d_{i(i+1)}$, $1 \leq i \leq (x-1)$ be the distance between the frames $\mathcal{F}_i$ and $\mathcal{F}_{i+1}$.
Let $d_{y(y+1)}$ be such that $d_{y(y+1)} \leq d_{i(i+1)}$, $1 \leq i \leq (x-1)$
Let $\mathcal{F}_{y'}$ be the average of the frames $\mathcal{F}_y$ and $\mathcal{F}_{y+1}$ (average of the corresponding elements of the two vectors corresponding to the two frames).
Then

$$\begin{aligned}\mathcal{N}^-(\mathcal{P}) &= \mathcal{N}^-(\{\mathcal{F}_1, \ldots, \mathcal{F}_y, \mathcal{F}_{y+1}, \ldots, \mathcal{F}_x\}) \\ &= \{\mathcal{F}_1, \ldots, \mathcal{F}_{y-1}, \mathcal{F}_{y'}, \mathcal{F}_{y+2}, \ldots, \mathcal{F}_x\}\end{aligned}$$

2. $\mathcal{N}^+ : \mathcal{S} \rightarrow \mathcal{S}$ is defined as follows: $\mathcal{N}^+(\mathcal{P}) = \mathcal{P}'$ if $\mid \mathcal{P}' \mid = \mid \mathcal{P} \mid + 1$.

We define $\mathcal{N}^+$ compositions r times as follows:
$(\mathcal{N}^+)^r(\mathcal{P}) = \mathcal{P}'$ if $\mid \mathcal{P}' \mid = \mid \mathcal{P} \mid + r$
When $r = 0$, we have $(\mathcal{N}^+)^0(\mathcal{P}) = \mathcal{P}$
Definition of $\mathcal{N}^+(\mathcal{P})$:
Let $\mathcal{P} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_x\}$ be a pattern.
Let $d_{i(i+1)}$, $1 \leq i \leq (x-1)$ be the distance between the frames $\mathcal{F}_i$ and $\mathcal{F}_{i+1}$.
Let $d_{y(y+1)}$ be such that $d_{y(y+1)} \geq d_{i(i+1)}$, $1 \leq i \leq (x-1)$
Let $\mathcal{F}_{y'}$ be the average of the frames $\mathcal{F}_y$ and $\mathcal{F}_{y+1}$
Then

$$\begin{aligned}\mathcal{N}^+(\mathcal{P}) &= \mathcal{N}^+(\{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_y, \mathcal{F}_{y+1}, \ldots, \mathcal{F}_x\}) \\ &= \{\mathcal{F}_1, \ldots, \mathcal{F}_y, \mathcal{F}_{y'}, \mathcal{F}_{y+1}, \ldots, \mathcal{F}_x\}\end{aligned}$$

$\mathcal{N}^-$ and $\mathcal{N}^+$ as defined above are clearly signal-dependent. In other words, frame reduction ($\mathcal{N}^-$) and frame expansion ($\mathcal{N}^+$) are based on signal characteristics. Signal-dependent analyses have been exploited by many researchers especially in the context of isolated word recognition [11].

In the following, we describe the various ways for defining a collection, $\mathcal{C}$, of reference patterns where $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_i, \ldots, \mathcal{C}_k\}$.

1. A Type I $\mathcal{C}_i$ is a collection of reference patterns corresponding to the i^{th} word in the vocabulary from multiple speakers.

2. A Type II $\mathcal{C}_i = \bigcup_{a=1}^{b} \mathcal{C}_{ia}$
where $\mathcal{C}_{ia}$ $(1 \leq a \leq b)$ is a singleton set that is defined as follows:
Let $\mathcal{C}_{ia'} = \{\mathcal{P}_{ia'j} \mid \mathcal{P}_{ia'j}$ is a pattern corresponding to the i^{th} word in the vocabulary from a speaker $\wedge$ $\mathcal{P}_{ia'j'} \in \mathcal{C}_{ia'} \Rightarrow \mid \mathcal{P}_{ia'j} \mid = \mid \mathcal{P}_{ia'j'} \mid\}$ for $1 \leq a' \leq b$
Then $\mathcal{C}_{ia} = \{\mathcal{P}_{ia1} \mid \mathcal{F}_{ia11} \in \mathcal{P}_{ia1}$ is the average of $(\mathcal{F}_{ia'11} \in \mathcal{P}_{ia'1}, \ldots, \mathcal{F}_{ia'|\mathcal{C}_{ia'}|1}), \ldots, \mathcal{F}_{ia1|\mathcal{P}_{ia'1}|}$ is the average of $(\mathcal{F}_{ia'1|\mathcal{P}_{ia'1}|}, \ldots, \mathcal{F}_{ia'|\mathcal{C}_{ia'}||\mathcal{P}_{ia'1}|})\}$

3. A Type III $\mathcal{C}_i$ is a singleton set that is defined as follows:
Let $\mathcal{C}_{i'} = \{\mathcal{P}_{i'j} \mid \mathcal{P}_{i'j}$ is a pattern corresponding to the i^{th} word in the vocabulary from a speaker$\}$
Let $\mathcal{P}_{i'x}$ be such that $\mid \mathcal{P}_{i'x} \mid \leq \mid \mathcal{P}_{i'y}$ for $1 \leq y \leq \mid \mathcal{C}_{i'} \mid$ *i.e.*, $\mathcal{P}_{i'y} \in \mathcal{C}_{i'}$
Let $\mathcal{C}_{i''} = \{\mathcal{P}_{i''j} \mid \mathcal{P}_{i'j} \in \mathcal{C}_{i'} \wedge (\mathcal{N}^-)^{|\mathcal{P}_{i'j}|-|\mathcal{P}_{i'x}|}(\mathcal{P}_{i'j}) = \mathcal{P}_{i''j}\}$
Then $\mathcal{C}_i = \{\mathcal{P}_{i1} \mid \mathcal{F}_{i11} \in \mathcal{P}_{i1}$ is the average of $(\mathcal{F}_{i''11} \in \mathcal{P}_{i''1}, \mathcal{F}_{i''21}, \ldots, \mathcal{F}_{i''|\mathcal{C}_{i''}|1}), \ldots, \mathcal{F}_{i1|\mathcal{P}_{i''1}|}$ is the average of $(\mathcal{F}_{i''1|\mathcal{P}_{i''1}|}, \ldots, \mathcal{F}_{i''|\mathcal{C}_{i''}||\mathcal{P}_{i''1}|})\}$

4. A Type IV $\mathcal{C}_i$ is a singleton set that is defined as follows:
Let $\mathcal{C}_{i'} = \{\mathcal{P}_{i'j} \mid \mathcal{P}_{i'j}$ is a pattern corresponding to the i^{th} word in the vocabulary from a speaker$\}$
Let $\mathcal{P}_{i'x}$ be such that $\mid \mathcal{P}_{i'x} \mid \geq \mid \mathcal{P}_{i'y} \mid$ for $1 \leq y \leq \mid \mathcal{C}_{i'} \mid$ *i.e.*, $\mathcal{P}_{i'y} \in \mathcal{C}_{i'}$
Let $\mathcal{C}_{i''} = \{\mathcal{P}_{i''j} \mid \mathcal{P}_{i'j} \in \mathcal{C}_{i'} \wedge (\mathcal{N}^+)^{|\mathcal{P}_{i'x}|-|\mathcal{P}_{i'j}|}(\mathcal{P}_{i'j}) = \mathcal{P}_{i''j}\}$
Then $\mathcal{C}_i = \{\mathcal{P}_{i1} \mid \mathcal{F}_{i11} \in \mathcal{P}_{i1}$ is the average of $(\mathcal{F}_{i''11} \in \mathcal{P}_{i''1}, \mathcal{F}_{i''21}, \ldots, \mathcal{F}_{i''|\mathcal{C}_{i''}|1}), \ldots, \mathcal{F}_{i1|\mathcal{P}_{i''1}|}$ is the average of $(\mathcal{F}_{i''1|\mathcal{P}_{i''1}|}, \ldots, \mathcal{F}_{i''|\mathcal{C}_{i''}||\mathcal{P}_{i''1}|})\}$

5. A Type V $\mathcal{C}_i$ is a singleton set that is defined as follows:
Let $\mathcal{C}_{i'} = \{\mathcal{P}_{i'j} \mid \mathcal{P}_{i'j}$ is a pattern corresponding to the i^{th} word in the vocabulary from a speaker$\}$
Let q be equal to $\lfloor(\mid \mathcal{P}_{i'1} \mid + \ldots + \mid \mathcal{P}_{i'm_{i'}} \mid)/m_{i'}\rfloor$ where $m_{i'}$ is $\mid \mathcal{C}_{i'} \mid$
Let $\mathcal{C}_{i''} = \{\mathcal{P}_{i''j} \mid \mathcal{P}_{i'j} \in \mathcal{C}_{i'} \wedge$
if $\mid \mathcal{P}_{i'j} \mid \geq q$ then $(\mathcal{N}^-)^{|\mathcal{P}_{i'j}|-q}(\mathcal{P}_{i'j}) = \mathcal{P}_{i''j}$
else $(\mathcal{N}^+)^{q-|\mathcal{P}_{i'j}|}(\mathcal{P}_{i'j}) = \mathcal{P}_{i''j}\}$
Then $\mathcal{C}_i = \{\mathcal{P}_{i1} \mid \mathcal{F}_{i11} \in \mathcal{P}_{i1}$ is the average of $(\mathcal{F}_{i''11} \in \mathcal{P}_{i''1}, \mathcal{F}_{i''21}, \ldots, \mathcal{F}_{i''|\mathcal{C}_{i''}|1}), \ldots, \mathcal{F}_{i1|\mathcal{P}_{i''1}|}$ is the average of $(\mathcal{F}_{i''1|\mathcal{P}_{i''1}|}, \ldots, \mathcal{F}_{i''|\mathcal{C}_{i''}||\mathcal{P}_{i''1}|})\}$

3 Algorithms for different methods

This section describes eight methods for comparing a test pattern $\mathcal{T}$ with a collection of reference patterns $\mathcal{C}$.

Method 1: The algorithm for this method is based on $\mathcal{A}^*$ algorithm [12]. The search space has $m * n$ nodes where m is the number of frames in the reference pattern and n is the number of frames in the test pattern. The start node is $(1,1)$ and the goal node is (m,n). Initially, the algorithm determines the nodes in the possible paths from the start node and chooses one with least distance. The alternatives along with distances are saved for later use. At any point, the algorithm expands that node which has least cumulative distance and this step is repeated till the goal

state is reached when the minimum achieved so far is this cumulative distance. The nodes (that have been saved) whose cumulative distance is greater than or equal to the current minimum distance are discarded thus pruning the search space. The algorithm terminates when there are no more paths to be explored.

Note that the above algorithm doesn't make use of window and slope constrains as described in [4]. Also observe that the high-cost paths are not explored fully.

Method 2: This method employs a matrix norm [13] to compare a test pattern with a collection of reference patterns. The matrix norm employed is the Frobenius norm.

Method 3 (Reduction): Let $\mathcal{T}$ be the test pattern and $\mathcal{P}_{ij}$ be a reference pattern. In this method the pattern ($\mathcal{T}$ or $\mathcal{P}_{ij}$) of larger length is reduced to the pattern ($\mathcal{T}$ or $\mathcal{P}_{ij}$) of smaller length with the help of $\mathcal{N}^-$ function.

Method 4 (Log-reduction): In order to investigate the non-linearity of the speech parameters, we employ in this method, log values of the parameters. This method is identical to Method 3 except that log values of the parameters are used.

Method 5 (Expansion): Let $\mathcal{T}$ be a test pattern and $\mathcal{P}_{ij}$ be a reference pattern. In this method the pattern ($\mathcal{T}$ or $\mathcal{P}_{ij}$) of smaller length is expanded to the pattern ($\mathcal{T}$ or $\mathcal{P}_{ij}$) of larger length with the help of $\mathcal{N}^+$ function.

Method 6 (Log-expansion): This method is identical to Method 5 except that log values of the parameters are used.

Method 7 (Reduction-cum-expansion): Let $\mathcal{T}$ be a test pattern and $\mathcal{P}_{ij}$ be a reference pattern. In this method both the patterns ($\mathcal{T}$ and $\mathcal{P}_{ij}$) are normalised to the average duration. Here, we make use of both $\mathcal{N}^+$ and $\mathcal{N}^-$ functions.

Method 8 (Log-reduction-cum-expansion): This method is identical to Method 7 except that log-values of the parameters are used.

4 Preliminary Results

4.1 Data Preparation

Speech data corresponding to the digits "one", "two", "six", and "seven" from ten female speakers forms the raw data for the generation of reference patterns. Here, a reference pattern is a sequence of frames where each frame is a vector of ten LPC coefficients extracted from speech data of 20 msecs duration. The uttered isolated digit is analyzed for beginning and ending silences and the thus sampled data is processed to create a sequence of ten LPC coefficients. For this purpose, we used *TI-Speech System*. Table 1 shows the details regarding the number of frames associated with various utterances.

Speaker	Frame Length for digits			
	one	two	six	seven
1.	6	5	8	8
2.	5	5	9	8
3.	6	5	8	10
4.	5	5	8	7
5.	8	8	9	7
6.	8	6	9	8
7.	7	8	12	7
8.	7	7	6	7
9.	7	7	4	8
10.	8	5	15	10

Table 1. Number of frames associated with utterances of "reference" speakers.

Thus, (from Table 2 (shown partially)), there are 40 Type I, 17 Type II, 4 Type III, 4 Type IV, and 4 Type V reference patterns.

Type	Class "one"	Class "two"
Type I	10	10
Type II	1(5)+1(6)+ 1(7)+1(8)	1(5)+1(6)+ 1(7)+1(8)
Type III	1(5)	1(5)
Type IV	1(8)	1(8)
Type V	1(6)	1(6)
Note: $n(m)$ denotes n patterns of m frames		

Table 2. Number of patterns for various of collection of reference patterns.

The "test" data consists of speech data corresponding to the digit "one" from three female speakers (different from ten speakers considered earlier). Table 3 provides summary of test data.

Speaker	Frame length for "one"
11	8
12	9
13	7

Table 3. Summary of test data.

4.2 Experiments

An experiment consists of using a particular *type* of collection of reference patterns and comparing a *test pattern* by employing a *method*. Thus there are $5 * 3 * 8 (= 120)$ experiments. The results of the experiments are displayed in the form 3-D plots for the purposes of comparison. One such plot is shown below. The 3-D plot shown has speakers along one dimension, methods along the second dimension, and normalized distances along the third dimension. The method dimension is further segmented and each segment consists of 4 units with each unit standing for a class. The distance normalization is carried out by determining the minimum and maximum distance between a test pattern and a collection of reference patterns. The minimum corresponds to the normalized 0 and the maximum corresponds to the normalized 100. The normalized distance for a class is obtained by averaging the normalized distances of the patterns in the class.

The ideal requirement would be to employ one of Type III, IV, or V collection of reference patterns (as

in these, the number of reference patterns in a class is minimum). It is observed from the plots that in these cases, Type V seems to provide better results (based on interclass and intraclass distances). For instance, in the plot corresponding to Method 5 and Speaker 1 in Plots 3, 4 and 5 (Plots 3 and 4 have not been provided), we observed that even though intraclass distances are small in all the three plots, the interclass distances with respect to the class "one" are more in Plot 5. Also, comparing Plot 1 and Plot 2 (again, both these plots have not been provided), we observed that results in Plot 2 are better. As the investigation is in a preliminary stage, we are not in a position to read too many things from the plots.

Acknowledgements

The first two authors would like to thank Dr. M. Narasimha Murty for helpful suggestions and useful discussions. The authors also thank Defence Research Development and Organization, India for the project grant DRDO/CSA/GK/13 that supported the research reported in this paper.

References

1. H. Wakita, "Normalization of vowels by vocaltract length and its application to vowel identification," *IEEE Trans. on ASSP*, Vol. 26, No. 2, pp. 183-192, April 1977.

2. H. Katterfelt,"A speaker-independent isolated word recogniser based on polynomial classifiers," *ICASSP-1988*, pp. 32-36, 1988.

3. R. Moore, "Computational techniques" in *Electronic speech recognition techniques, technology & applications*, G. Bristow(Editor), McGraw-Hill company, 1986.

4. H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on ASSP*, Vol. 26, No. 1, February 1978.

5. K. K. Paliwal, A. Agarwal and S. S. Sinha, "A modification over Sakoe and Chiba's DTW algorithm for isolated word recognition," *Signal Processing*, Vol. 4, No. 4, pp. 329-333, 1982.

6. S. Haltsonen, "Improved DTW methods for discrete utterance recognition," *IEEE Trans. on ASSP*, Vol. 33, No. 2, April 1985.

7. R. Pieraccini, "Pattern compression in isolated word recognition ," *Signal Processing*, Vol. 7, No. 1, pp. 1-15, 1984.

8. M. H. Kuhn, H. Tomaschewski and H. Ney, "Fast nonlinear time alignment for isolated word recognition," *ICASSP-1981*, pp. 736-740, 1981.

9. V. R. Viswanathan, J. Makhoul, R. M. Schwartz and A. W. F. Huggins, "Variable frame rate transmission: a review of methodology and application to narrow-band LPC speech coding,"*IEEE Trans. on Communications*, Vol. 30, No. 4, pp. 674-686, 1982.

10. C. K. Chuang and S. Chan, "Speech recognition using variable frame rate coding," *ICASSP-1983*, pp. 1033-1036, 1983.

11. B. Yegnanarayana and T. Sreekumar, "Signal-dependent matching for isolated word speech recognition systems," *Signal processing*, Vol. 7, No. 2, pp. 161-173, 1984.

12. E. Rich, *Artificial intelligence*, McGraw-Hill company, 1983.

13. G. W. Stewart, *Introduction to matrix computations*, Academic Press, 1973.