International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:4, 2014

Bidirectional Dynamic Time Warping Algorithm for the Recognition of Isolated Words Impacted by Transient Noise Pulses

G. Tamulevičius, A. Serackis, T. Sledevič, D. Navakauskas

***Abstract*—**We consider the biggest challenge in speech recognition – noise reduction. Traditionally detected transient noise pulses are removed with the corrupted speech using pulse models. In this paper we propose to cope with the problem directly in Dynamic Time Warping domain. Bidirectional Dynamic Time Warping algorithm for the recognition of isolated words impacted by transient noise pulses is proposed. It uses simple transient noise pulse detector, employs bidirectional computation of dynamic time warping and directly manipulates with warping results. Experimental investigation with several alternative solutions confirms effectiveness of the proposed algorithm in the reduction of impact of noise on recognition process – 3.9% increase of the noisy speech recognition is achieved.

***Keywords*—**Transient noise pulses, noise reduction, dynamic time warping, speech recognition.

I. Introduction

NOISE reduction is one of the biggest challenges in a speech recognition task. At laboratory conditions attained high speech recognition rate in real-world tasks will be downgraded by various types of background noise, other speech sources, channel distortion, speech variability [1]–[4].

Transient noise pulses are generated in household and office environment by various physical processes (knocking or closing the doors, switching of nearby electric devices, keyboard typing, etc.). The transient noise pulses appear randomly and it is impossible to use conventional noise reduction methods like spectral subtraction or cepstral mean subtraction to eliminate them.

In this paper we concentrate on transient noise pulses elimination task in order to improve downgraded speech recognition rate. We propose the bidirectional dynamic time warping (DTW) algorithm helping to reduce transient noise impact on speech recognition process. The proposed bidirectional DTW algorithm evaluates similarity of speech utterances in noise uncorrupted segments thus reducing the impact of noise on recognition results.

II. Transient Noise Pulse Removal

Two main approaches can be formulated for the improvement of noisy speech recognition. The first one is the enhancement of speech data (testing data) ant the second one is the adaptation of the model (training data).

The speech data can be enhanced in two ways – processing the speech signal itself or enhancing the extracted features. Microphone array based beam forming amplification [5], various filters like median, noise suppression, nonlocal neighborhood filters [6], spectral subtraction based enhancement approaches [7] are used to reduce the mismatch between training and testing acoustical conditions. This can be complicated considering the redundancy of the speech signal, non-stationarity and variable level of transient noise pulses. An alternate approach aims to extract robust features or to enhance extracted recognition features thus making speech recognition robust to noise. Various perceptually motivated feature systems are proposed as to some degree noise robust, e. g.: perceptual linear prediction analysis [8], zero-crossings with peak amplitude analysis [9], gammatone frequency cepstral analysis [10], power-normalized cepstral analysis [11]. The aim of the feature enhancement procedure is the modification of features into more noise robust form. There are various enhancement techniques proposed for robust speech recognition, too: cepstral mean normalization, cepstral mean subtraction, RASTA filtering, and other techniques like minimum-mean-square-error noise reduction [12] or bias-residual decomposition of features [13].

The idea of the model adaptation is to modify acoustic models considering the noise environment. This approach has the advantage over the speech enhancement approaches but is more computationally complicated. The predefined nature of the approach causes its incapability to process varying in time noise process like transient noise pulses.

Transient noise pulses appear like sharply rising impulses with following transient part which can last a few hundred milliseconds. The fading part of the noise is a result of resonance processes. As it is non-stationary process the most effective way to discard its impact on recognition process is to discard the transient pulse noise. The removal of transient noise pulses uses the model of the pulse. It can be modeled using temporal template model, linear predictive model or hidden Markov model [14]. After removal of the corrupted segment the missing segment of the signal can be restored using linear prediction modeling, interpolation [14], [4].

G. Tamulevičius is with the Vilnius University Institute of Mathematics and Informatics, Akademijos str. 4, LT-08663 Vilnius, Lithuania (e-mail: gintautas.tamulevicius@mii.vu.lt).

A. Serackis, T. Sledevič and D. Navakauskasare with the Department of Electronic Systems, Vilnius Gediminas Technical University, Naugarduko str. 41-413, LT-03227 Vilnius, Lithuania.

III. Transient Noise Pulse Detection

The removing of frames with corrupted speech will be performed in two steps: detection of the transient pulse in a speech signal and the removal of corrupted segment.

Exceptional attribute of the transient pulse is a sharp rise which gives abrupt change of the signal properties thus making it possible to detect by modeling the speech.

Our proposed method is based on assumption that speech can be modeled using linear prediction whereas the transient pulse cannot. Thus the increase of prediction error at the moment of transient pulse beginning could be expected.

The increase of the prediction error can be detected as the rise of the prediction error. The change rate of the function is defined as the time derivative of the function. In our work we expressed the derivative of the prediction error as follows

$$\frac{\partial e(t)}{\partial t} \approx \frac{\Delta e(n)}{\Delta n} = \frac{e(n)-e(\sigma)}{\sigma} = e(n)-e(n-1), \qquad (1)$$

here $e(n)$ – prediction error, n – number of frames, σ– error calculation interval (σ = 1 in our case).

Frame with a sharp rise of the prediction error can be denoted as the starting frame of the transient pulse (Fig. 1).

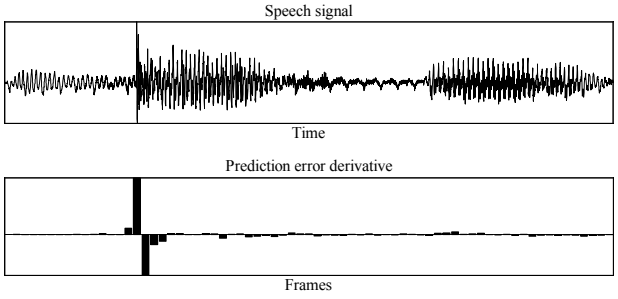


Fig. 1 Speech signal and its prediction error rise

Another decision is the number of frames to denote as corrupted. Again, we used the approximation of the time derivative of the prediction error. The drop of the prediction error change rate will mark the end of the distorted speech segment. For this purpose we used the predefined threshold value to determine the end of the noise segment.

After the noise segment is detected the second step is removal of the segment. We assume that discarding of a few corrupted speech frames with the length of 20–80ms can reduce the noise impact on recognition rate. Despite the fact that speech signal is removed too, we assume that rejection of noise distorted data will yield higher impact than loss of overlapped speech data.

Next we formulate the modified dynamic time warping algorithm allowing us to reduce the impact of the transient pulse noise on word recognition process thus improving the recognition robustness.

IV. Dynamic Time Warping Algorithm Modification

A. Dynamic Time Warping Algorithm

The Dynamic Time Warping (DTW) is a pattern matching algorithm for comparison of time sequences and is widely used for speech recognition, string and character recognition, data mining tasks. Traditionally DTW is implemented using Dynamic programming principle [15].

The main advantages of DTW algorithm are its simplicity, capability to compare sequences of different length, and independence of comparison unit.

The point of DTW algorithm is to match two time sequences and to calculate their distance in most coinciding points. The calculated distance can be used for making decision on similarity of compared sequences. DTW algorithm can be presented as the search of the minimal cost path in the grid (Fig. 2 (a)).

Every point of the grid represents the pair of compared feature vectors and carries numerical value of distance between vectors. During the sequences comparison these points are analyzed in a search of path through the grid with minimal accumulated distance. Minimal accumulated (overall) distance implies maximal similarity of sequences.

In order keep the search meaningful for time sequences there are used some restrictions for search procedure. They form so called warping function which controls the search of the path. Generally the following restrictions are used [16]:

- Endpoint conditions – start the search in the point (l, M) and terminate the search in point (M, N);
- Step size is limited to one vertical, horizontal or diagonal movement between points;
- Global constraints restrict the search area (see the grey zones in Fig. 2) and exclude meaningless comparison.

The overall distance of the compared sequences is expressed

$$D = \min \sum_{n=1}^{N} \varphi\left[d\left(f_n, f_m \right) \right], \qquad (2)$$

here N and M are lengths of the compared sequences, φ[·] – warping function, f_n – n-th feature vector of the sequence.

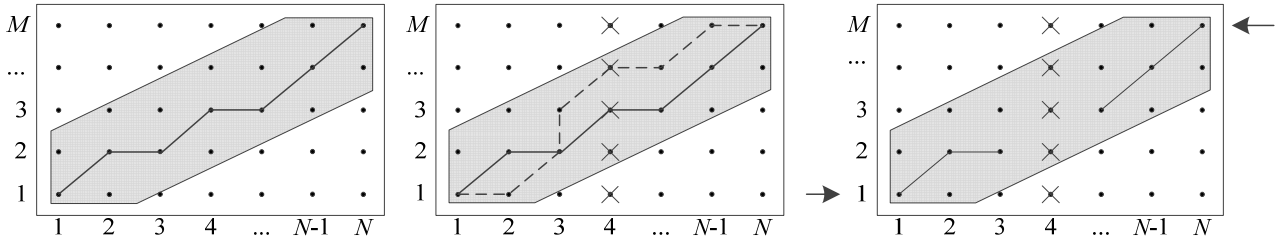
B. Bidirectional Dynamic Time Warping Algorithm

Comparing sequences A and B we get the overall distance calculated in most coinciding points. Traditionally we start our search in the point (1, 1) and finish in (N, M) as the endpoint conditions claim. The traced path with minimal distance is the only possible as the calculated distance is optimal globally. And this path is determined in the last point of search because only in the point (N, M) we get the overall minimal distance. Only then we can trace the path.

If requirements of symmetry are fulfilled, i. e., if we use symmetrical form of local and global constraints, we should get the identical overall distance despite of the search direction (Fig. 2 (a)). Therefore, if we start the search of minimal distance in the point (N, M) and finish in (1, 1) we will get the same minimized overall distance.

Comparing A with noisy version of B (the particular segment of B is corrupted only) we will get different value of overall distance and another path in search grid (Fig. 2 (b)).

The reason for this will be distorted features of the corrupted segment and their impact on global path and overall distance.



(a) noise free unidirectional search (b) unidirectional search on noisy data (c) bidirectional search on noisy data

Fig. 2 Dynamic time warping based search

Aiming to avoid of distorted feature impact on overall distance we should remove them from comparison process. This could be done in various manners.

The simplest way is to remove the noise corrupted segment of the signal thus discarding distorted features. Formally we will fasten the comparison process (reduce the amount of the acoustical data) and remove the noise from the distance calculation. However, the removal of corrupted speech segment with acoustical noise will result in another distortion in signal clipping region. And this distortion can condition another modification of global path.

Another analyzed noise removal procedure was the discarding of distorted feature effect on global distance. In this case the search of global minimum path is performed using the whole set of features (distorted even) but the global distance is calculated ignoring them. The main assumption for this procedure was that distorted features impact all comparison processes and their impact more or less comparable for all comparison cases.

Our main proposal is the bidirectional DTW algorithm. The idea is to perform DTW-based comparison starting from both ends of sequences. The main assumptions for this were:

- The distortion appears in random sequence place. The range of distorted signal is known;
- Partially calculated global distance does not reveal the global distance or path. It reflects similarity of compared parts only;
- Comparison can be performed from both sides of sequences. In case of identical sequences comparison will give identical result. In case of different sequences we will get similarity of starting and ending segments.

According to these assumptions the comparison of sequences A and distorted version of sequence B will consist of two steps and will be as follows:

- Firstly the comparison started from the beginning of the sequence to the beginning of the corrupted segmented of sequence B (see Fig. 2 (c)). We obtain the partial distance between starting segments of the sequences (the starting segment of sequence A and the uncorrupted segment of sequence B);
- The second comparison is started at the end of the sequences and continued till the end of the distorted segment of the sequence B. Thus we obtain the partial distance of ending clean segments;
- The overall similarity of sequences is calculated by sum of partial distances

$$D = D_1 + D_2.$$

We entitle this modification of DTW algorithm as bidirectional dynamic time warping algorithm.

In order to get partial distances closer to overall distance of one directional DTW algorithm as much as possible we will use the same local and global constraints for the comparison procedure.

We state that overall distance D calculated using bidirectional DTW algorithm will represent the similarity of sequence A and coincident clean segments of sequence B. And this distance can be used for speech pattern classification task. We will test our formulated bidirectional comparison algorithm in experimental study of noise corrupted speech recognition.

V. Experimental Results

Experimental analysis of the proposed approach was performed. The removal of corrupted speech frames was applied for isolated word recognition task. For this purpose 8 speakers (4 males and 4 females) pronounced 100 different words two times.

The first session records were used for training. The second session records were used for testing. Two versions of these records were created: clean and noise corrupted. The latter was created by adding transient pulse noise artificially. 100–200 ms length transient pulse segments with random scale factor were overlapped with random segments of recorded words. The records of closing drawer, pencil knocking the table and similar sounds were used for this purpose. The signal-to-noise ratio (SNR) of the noise interrupted records varied from −6 dB to −12 dB.

The DTW-based isolated word recognition system was used

for experimental research. The 12th order linear prediction coding cepstral analysis (LPCC) was used for feature extraction [17], [18].

The goal of the first experiment was to evaluate the recognition rates for clean and noisy speech records. The results of the experiment are given in Table I.

The individual recognition rate of clean speech varied from 90 % to 99 % giving the average rate of 96.6 %.

As we can see addition of transient pulse noise reduced the recognition rate by 12 % approximately. In some cases rate decrease exceeded 20 % for particular speaker.

TABLE I
RECOGNITION RATES FOR CLEAN AND NOISY SPEECH

Speaker	Clean speech	Noisy speech
Speaker 1	99 %	97 %
Speaker 2	99 %	95 %
Speaker 3	98 %	93 %
Speaker 4	99 %	82 %
Speaker 5	94 %	83 %
Speaker 6	97 %	76 %
Speaker 7	90 %	71 %
Speaker 8	97 %	78 %
Average	96.6 %	84.4 %

The second experiment was intended for evaluation of noise corrupted speech frames removal impact on recognition rate.

Three different corrupted speech discarding procedures were applied:

- P1 – removing the noise corrupted segment of the speech signal;
- P2 – discarding the distances of the distorted features from the distance calculation process;
- P3 – bidirectional DTW-based comparison of the words.

The records of the first session were used as reference (training) data; the records of the second session with added noise were used as test data. The results are presented in Table II.

TABLE II
RECOGNITION RATES FOR DIFFERENT DISCARDING PROCEDURES

Speaker	P1	P2	P3
Speaker 1	97 %	96 %	94 %
Speaker 2	96 %	95 %	93 %
Speaker 3	93 %	93 %	93 %
Speaker 4	85 %	84 %	87 %
Speaker 5	84 %	86 %	89 %
Speaker 6	77 %	80 %	85 %
Speaker 7	72 %	73 %	78 %
Speaker 8	80 %	81 %	87 %
Average	85.5 %	86 %	88.3 %

As we could expect there is no any big difference between the removing noise corrupted speech segment and the discarding of distorted features from distance calculation process, the difference of rates was smaller than 1%.

Analysis of the results has shown that the rate of correct detection of noise corrupted segment starting moment was 98.4. The length of detected and discarded noisy segments varied from 20ms to 80ms. Considering the average length of tested speech utterances the comparison process was accelerated up to 10%.

VI. CONCLUSION

The bidirectional DTW algorithm was proposed for comparison of sequences with noise insertions. The comparison is performed with assumption about known region of noise in speech and gives similarity of compared clean regions of speech utterances.

The proposed algorithm was tested experimentally. The bidirectional DTW-based comparison improved recognition rate of noisy speech by 3.9% and overtook other noisy signal discarding procedures by 2.3%.

Experiments revealed that longer segments were discarded from records with lower SNR value. Hence there exists relationship between the recording quality, the length of discarded segment and its impact on recognition process. The higher quality we have, the shorter segment we need to discard in order to improve recognition process.

ACKNOWLEDGMENT

This research was funded by a grant (No. MIP-092/2012) from the Research Council of Lithuania.

REFERENCES

[1] L. Deng, X. Huang, "Challenges in adopting speech recognition", *Communications of ACM*, vol. 47(1), ACM, New York, pp. 69–75, 2004.
[2] G. Čeidaitė, L. Telksnys, "Analysis of factors influencing accuracy of speech recognition", *Electronics and Electrical Engineering*, no. 9(105), pp. 69–72, 2010.
[3] Ch.-P. Chen, *Noise robustness in automatic speech recognition*, Ph. D. thesis, University of Washington, 2004.
[4] J. Benesty, S. Makino, J. Chen, *Speech enhancement*, Berlin: Springer-Verlag, 2005.
[5] M. Seltzer, M. *Microphone array processing for robust speech recognition*, Ph. D. thesis, Carnegie Mellon University, Pittsburgh, 2003.
[6] R. Talmon, I. Cohen, and Sh. Gannot, "Transient noise reduction using nonlocal diffusion filters", *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19(6), pp. 1584 – 1599, 2011.
[7] R. Gomez and T. Kawahara, "Optimizing spectral subtraction and Wiener filtering for robust speech recognition in reverberant and noisy conditions", in *Proc. of ICASSP*, pp. 4566–4569, 2010.
[8] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *Journal of Acoustical Society of America*, vol. 87(4), pp. 1738–1752, 1990.
[9] D.-S. Kim, S.-Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments", *IEEE Trans.Speech and Audio Processing*, vol. 7(1), pp. 55–69, 1999.
[10] Y. Shao, Zh. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition", in *Proc. of ICASSP*, pp. 4625–4628, 2009.
[11] Ch. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction", in *INTERSPEECH 2010*, pp. 2058–2061, 2010.
[12] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, „A Minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition", in *Proc. of ICASSP*, pp. 4041–4044, 2008.
[13] M. Fujimoto, S. Watanabe, and T. Nakatani, "Non-stationary noise estimation method on bias-residual component decomposition for robust speech recognition", in *Proc. of ICASSP*, pp. 4816–4819, 2011.

[14] S.V. Vaseghi, *Advanced digital signal processing and noise reduction*, New York: Wiley, 2006.
[15] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans.Speech and Audio Processing*, vol. 26(1), pp. 43–49, 1978.
[16] L. Rabiner and B.-H. *Juang Fundamentals of speech recognition*, New Jersey: Prentice-Hall, 1993.
[17] T. Sledevič, D. Navakauskas, "FPGA based fast Lithuanian isolated word recognition system", in *Proc. of EUROCON 2012*, pp. 1630–1636, 2013.
[18] T. Sledevič, G. Tamulevičius, D. Navakauskas, "Upgrading FPGA implementation of isolated word recognition system for a real-time operation", *Electronics and Electrical Engineering*, no. 10(19), pp. 123–128, 2013.