Speech Recognition of a Limited Vocabulary Using the Convolution Kernel Compensation Approach

DAMJAN ZAZULA, GREGOR KREBS
Faculty of Electrical Engineering and Computer Science
University of Maribor
Smetanova 17, 2000 Maribor
SLOVENIA
http://storm.uni-mb.si

Abstract: This paper introduces a simple speech recognition approach based on the convolution kernel compensation (CKC). The processing method is reveal in detail and applied to a specific vocabulary of 11 Slovenian words suitable to control a wheelchair. Experimental results are compared with the outcomes of two more sophisticated approaches, i.e. dynamic time warping (DTW) and neural networks (NN). The obtained recognition rates are 67.12 % for NN, 90 % for our CKC-based method and 97.72 % for DTW.

Key-Words: Speech recognition, Limited vocabulary, BSS, Convolution kernel compensation, Neural networks

1 Introduction

Speech processing is one of the most propulsive research fields of today. A variety of recognition and synthesis algorithms have been proposed. Recently, most of them try to solve problems of man-machine communication and automated language interpretation in as natural and thorough way as possible.

An automated speech recognition procedure starts with the acoustic signal recording and preprocessing. In this stage, it is important to eliminate most of the background and measurement noise as well as all possible artefacts. Usually, a proper signal pre-conditioning is achieved by the so-called voice activity detection (VAD) algorithms which look for the voiced segments of speech, i.e. more or less individual spoken words [1]. A well-known fact is that the speech signals show a lot of nonstationarity. This prevents a recognition based on the entire speech signal segments. The most spread technique to cope with speech nonstationarities is construction of Mel-frequency cepstral coefficients (MFCC) [2]. Computation of cepstral coefficients separates the transfer function of vocal tract from the excitations. Actually, the excitation is eliminated from further processing. As this is done by using a bank of Mel-frequency filters, the vocal tract transfer function, depicted by the cepstral coefficients, is featured at those frequencies only which are most characteristic in human speech.

The obtained speech features do not correspond directly to, say, individual phonemes. They merely represent successive units of the analysed speech, so the sequences of those units form chains of speech states. Transition probabilities between the states comply with the so-called hidden Markov models [3].

Apart from the problem of general speech recognition, there are quite frequent situations where a limited vocabulary of just a few words satisfies. Imagine, for example, ordering goods from a slot-machine or even verbally controlling a car drive. Very similar vocabulary can also be applied in the case of a speech-controlled wheelchair. The students in electrical engineering and computer science at the Faculty of Electrical Engineering and Computer Science, University of Maribor in Slovenia, completed such a wheelchair by successfully using a neural-network-based speech recognition algorithm capable of recognizing 11 Slovenian words [4].

Along with the wheelchair project, we also tested the efficacy of some widely used speech recognition methods experimenting with those 11 Slovenian words [5]. In parallel, we preliminary verified a simple statistical approach based on the convolution kernel compensation (CKC). This paper compares the results obtained by CKC with two other approaches, i.e. with dynamic time warping (DTW) and neural networks (NN). In Section 2, the basics of a novel, CKC-based source separation technique is explained. Section 3 introduces a CKC-based speech recognition algorithm for limited vocabularies, while Section 4 uses it for the recognition of a corpus of 11 Slovenian words and compares the results with those obtained by two abovementioned more sophisticated methods. The paper is concluded with a discussion in Section 5.

2 CKC-Based Source Separation

Recapitulate briefly the CKC basics as implemented in [6] for pulse source reconstruction and separation. Consider the following data model:

$$\mathbf{x}(n) = \mathbf{H}\mathbf{s}(n) + \mathbf{v}(n) \tag{1}$$

where:

$\mathbf{x}(n) = [x_1(n),...,x_M(n)]^T$ is a vector of M observations;

$\mathbf{s}(n) = [s_1(n),...,s_K(n)]^T$ describes a vector of K sources which are mutually independent over certain period of time;

$\mathbf{H}$ is an $M \times K$ mixing matrix which stands for the convolutive relationship but is otherwise unknown;

$\mathbf{v}(n) = [v_1(n),...,v_M(n)]^T$ is an i.i.d noise vector independent from the sources.

To extend relationship (1) from convolutive to a multiplicative MIMO vector form, the vector $\mathbf{x}(n)$ has to be augmented by M_e delayed repetitions of each observation:

$$\mathbf{x}_e(n) = [x_1(n),....,x_1(n-M_e+1),....,x_M(n),....,\\ x_M(n-M_e+1)]^T \quad (2)$$

where M_e is assumed to satisfy

$$M \cdot M_e \geq K(L+M_e), \quad (3)$$

and L stands for the length of the transmission channel responses, i.e. the constituent signal components.

Extending the noise vector in the same manner, (1) can be rewritten in a vector form:

$$\mathbf{x}_e(n) = \mathbf{H}_e \mathbf{s}_e(n) + \mathbf{v}_e(n). \quad (4)$$

$\mathbf{H}_e$ in (4) contains the channel unit sample responses $h_{ij}(l)$:

$$\mathbf{H}_e = \begin{bmatrix} \mathbf{H}_{11} & \cdots & \mathbf{H}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{M1} & \cdots & \mathbf{H}_{MK} \end{bmatrix} \quad (5)$$

with

$$\mathbf{H}_{ij} = \begin{bmatrix} h_{ij}(0) & \cdots & h_{ij}(L) & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & h_{ij}(0) & \cdots & h_{ij}(L) \end{bmatrix}, \quad (6)$$

while the extended vector of sources $\mathbf{s}_e(n)$ takes the following form:

$$\mathbf{s}_e(n) = [s_1(n),....,s_1(n-L-M_e+1),...,s_K(n),...,\\ s_K(n-L-M_e+1)]^T \quad (7)$$

The correlation matrix of the extended observations can be computed as:

$$\mathbf{R}_{x_e} = \lim_{T\to\infty} \frac{1}{T} \sum_{n=1}^{T} \mathbf{x}_e(n)\mathbf{x}_e^*(n) = \mathbf{H}\mathbf{R}_{\mathbf{s}_e}\mathbf{H}^T + \sigma^2\mathbf{I} \quad (8)$$

where $\mathbf{R}_{\mathbf{s}_e}$ denotes the correlation matrix of sources and $\mathbf{x}_e^*(n)$ stands for the conjugate transpose of $\mathbf{x}_e(n)$.

2.1 Activity Index

If the transmission channel responses, i.e. the signal components, differ, matrix $\mathbf{H}$ has full column rank $rank(\mathbf{H}) = K(L+M_e)$. Then, a so-called activity index can be introduced (calculated as Mahalanobius distance):

$$Ind(n) = \mathbf{x}_e^T(n)\overline{\mathbf{R}}_{\mathbf{x}_e}^{-1}\mathbf{x}_e(n) = \\ \mathbf{s}_e^T(n)\mathbf{H}^T(\mathbf{H}\mathbf{R}_{\mathbf{s}_e}\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{s}_e(n) + v_{\mathbf{n}_e}(n) = \\ \mathbf{s}_e^T(n)\mathbf{H}^T(\mathbf{H}^T)^{-1}\mathbf{R}_{\mathbf{s}_e}^{-1}\mathbf{H}^{-1}\mathbf{H}\mathbf{s}_e(n) + v_{\mathbf{n}_e}(n) = \\ \mathbf{s}_e^T(n)\mathbf{R}_{\mathbf{s}_e}^{-1}\mathbf{s}_e(n) + v_{\mathbf{n}_e}(n), \quad (10)$$

where superscript $^{-1}$ denotes the matrix inverse and $v_{\mathbf{n}_e}(n)$ replaces the impact of all noise. In the noise-free case, activity index *Ind(n)* differs from zero only at those time instants where at least one source is active. Its value is proportional to the number of simultaneously active sources.

2.2 Noise-Free Decomposition

Suppose only the *i*-th source active at the time instant n_0. According to (11), the entire pulse train of this source can be reconstructed as

$$p_{n_0,i}(n) = \mathbf{x}_e^T(n_0)\overline{\mathbf{R}}_{\mathbf{x}_e}^{-1}\mathbf{x}_e(n) = \\ \mathbf{s}_e^T(n_0)\mathbf{R}_{\mathbf{s}_e}^{-1}\mathbf{s}_e(n) = r_{i,i}s_{e,i}(n_0)s_{e,i}(n) \quad (11)$$

where $r_{i,j}$ denotes the *(i,j)*-th element of $\mathbf{R}_{\mathbf{s}_e}^{-1}$.

Once the instants of appearance of the signal components, i.e. the source pulse trains, are known, also the components themselves can be obtained from the given observations–for example, by using the spike-triggered averaging approach.

3 CKC-Based Speech Recognition

In the previous section, we explained the basics of the CKC approach for blind source separation. It is base d on a MIMO model which supposes that the number of observations exceeds the number of sources. If this is true, the correlation matrix of observations has full column rank and, thus, it leads to a thorough source separation. Dealing with an underdetermined system, no ideal source separation is guaranteed any more. The activity index (Eq. (10)) and the decomposed pulse trains (Eq. (11)) do not depend only on the sources' activities any more, because the influence of an incompletely compensated convolution kernel degrades their features. Empirical findings from [6] allude that a thorough decomposition is still possible as long as the number of sources stays below twice the number of observations. On the other hand, an upgraded nonlinear approach from [7] proves that even with a single observation, i.e. considering a MISO model, one can count on a successful blind source separation.

In the majority of cases, speech signals are recorded with one microphone only. This means that just one observation is available. If the speech were stationary the decomposition method from [7] could be applied. Because it is not so, we tried to make use of the signal's

compound character and split its voiced segments into several intervals of, presumably, stationary activity. We followed the findings published elsewhere on the length of stationary epochs being about 25 to 30 *ms*. Considering such signal intervals as fundamental and independent components, the most natural MISO interpretation can be transformed in a MIMO model. This new, rather artificial point of view supposes a speech segment integrants, i.e. a sequence of stationary epochs, become the model outputs appearing at the same moment. Of course, the role of the input source excitations also changes in that the excitations of given consecutive stationary epochs become parallel and simultaneous. In the sequel, we are going to encompass this new image with an appropriate data model.

First, assume in Eq. (1) only one noise-free observation. Consequently, the convolution kernel **H** shrinks into a convolution vector $\mathbf{h}_1$:

$$x_1(n) = \mathbf{h}_1\mathbf{s}(n);\ n = 0,...,N-1 \tag{12}$$

Now, split this observation into M parts of equal length:

$$y_i(n) = x_1([i-1]\frac{N}{M} + n);\ i = 1,\ldots,M,\ n = 0,\ldots,\frac{N}{M}-1 \tag{13}$$

To reformulate (12) accordingly, we have to think about a modification of $\mathbf{h}_1$ and $\mathbf{s}(n)$. The input sources $s_j(n)$; $j = 1,\ldots,K$, as already defined in (1), reduce in number because all the sources that originally trigger at a time instant $n = n_0 + (i-1)\frac{N}{M};\ i = 1,\ldots,M$, now concentrate in a single source which triggers at n_0.

Denote these combined sources by $u_j(n)$; $j = 1,\ldots,K_m$, where K_m stands for an unknown number of modified sources. On the other hand, the convolution vector $\mathbf{h}_1$ combines into M vectors. Each of them comprises the contributions of all the system channel responses (due to the original MISO interpretation) that build up the corresponding segment of the original observation. So, $y_i(n_0)$ needs all the samples from $\mathbf{h}_1$ which, according to $\mathbf{s}(n_0)$, sum up in this very moment, n_0. Denote this newly obtained convolution kernel of dimensions $M \times K_m$ by $\mathbf{H}_\mathrm{m}$. Hence, we are back to a MIMO model which, in a noise-free case, yields:

$$\mathbf{y}(n) = \mathbf{H}_m\mathbf{u}(n);\ n = 0,\ldots,\frac{N}{M}-1 \tag{14}$$

Eq. (14) may be considered formally the same as (1), therefore the decomposition hints derived in Section 2 up to Eqs. (10) and (11) hold analogously in this new situation.

How can $\mathbf{H}_\mathrm{m}$ and $\mathbf{u}(n)$ be interpreted from a standpoint of speech segments, i.e. presumably spoken words? By splitting the recorded speech into intervals of a length characteristic for the stationary epochs, one can consider the rows in $\mathbf{H}_\mathrm{m}$ contain the corresponding stationary speech components. At the same time, $\mathbf{u}(n)$ stands for the unified artificial sources that trigger all the speech components present at n in $\mathbf{y}(n)$.

If the splitting applied in Eq. (13) generates such a constellation that $M > K_m$, then a thorough decomposition may be foreseen using Eqs. (10) and (11). This decomposition is going to separate the artificial sources $\mathbf{u}(n)$, whereas the convolution kernel $\mathbf{H}_\mathrm{m}$ would be compensated and, thus, its influence eliminated. However, a proper extension factor M_e must be found referring to Eq. (3).

Taking all this into account, how can a speech recognition procedure benefit out of it? Recall the initial assumption on a limited vocabulary. For every known reference word from this vocabulary a correlation matrix $\mathbf{R}_\mathrm{ye}$ can be constructed by combining the derivations from Eqs. (14) and (8). Moreover, a learning set of speech segments belonging to different classes of words may be used in such a way that several words from the same class, say the l-th one, contribute to the same correlation matrix, say $\mathbf{R}_\mathrm{ye}^{(l)}$. Using Eq. (10), activity indexes can be computed for any unknown word with all $\mathbf{R}_\mathrm{ye}^{(l)}$: $Ind^{(l)}(n);\ n = 0,\ldots,\frac{N}{M}-1,\ l = 1,\ldots,\Lambda$, where Λ denotes the number of reference words.

Empirically, we found out that the mean values of activity index $Ind^{(l)}(n)$; $l = 1,\ldots,\Lambda$, calculated by (10), exhibit a minimum at that l which corresponds to the correct reference for an unknown word. Owing to a rather high variability of recorded speech in real environment, this measure can sometimes give false positives with references which are also very close to the correct one. Actually, the mean values of activity indexes can sometimes be very close together. In such cases, we introduce another distinctive measure. Using all the reference correlation matrices whose indexes for an unknown word are close and under a preselected threshold T_m, also the pulse trains according to (1) are computed for them. Position n_0 which extracts a pulse train is taken as a minimum-value sample index in the activity index with the lowest mean among all the compared activity indexes. Excluding the sample at position n_0, the pulse train with the lowest variability (the smoothest one) is supposed to indicate the most thorough kernel compensation. And further, the convolution kernel of an unknown word is expected to be optimally compensated exactly by the correct reference. So, the smoothest pulse train, to our experience, sorts out the best kernel compensating reference, which means the final decision in our word recognition approach.

3.1. Computational Algorithm

The proposed CKC-based speech recognition method is compacted in the following computational steps.

I. Construction of a base of reference words

1. Obtain a single measurement of a speech signal.
2. Filter out the most representative speech frequency spectrum (approx. 500 to 1500 *Hz*).
3. Apply a VAD search and extract only active parts of the signal (presumably separate spoken words).
4. Truncate the obtained signal segments to a preselected length N (say, of duration of 1 *s*) and concatenate them into a learning vector $x_l(n)$.
5. Build a MIMO model using Eqs. (13) and (14)– mote that the segment length, N, must be a divisible by the number of artificial observations, M, otherwise the segments must be padded by zeros to an appropriate length.
6. Now, calculate the correlation matrices according to Eq. (8) for all the reference words (known speech segments) and save them for the recognition purposes.

II. Recognition of unknown speech segments

1. Record an unknown speech segment and preprocess it the same way as in the points I.2, I.3, and I.4 in that part talking about the truncation.
2. Now, we have a speech segment of length N. First, calculate its activity indexes according to all known reference words (Eq. (10)). Leave out the starting and ending transition intervals of length K.
3. If the mean values of several indexes differ for less than T_m, compute their pulse trains (Eq. (11)). The extraction position n_0 is taken to be at the lowest value of the minimum-mean activity index.
4. Recognise the word:
   - if point II.3 was applied, the pulse train with minimum variance decides the correct reference word;
   - if point II.3 was skipped, the minimum-mean activity index decides the correct reference word.

4 Experimental results

We experimented with a vocabulary of 11 Slovenian words, all spoken by the same male person in a real room environment, but with all acoustic disturbances kept as low as possible. The recordings were done by a non-professional, OC multimedia microphone, sampled with 11 *kHz* and 16-bit resolution.

The recognition results of three different approaches will be compared in this paper. Two of them are well-known methods based on DTW [8] and neural networks [9]. Because of limited space, we are not going to describe their implementation details which are revealed in [5]. In the first place, we are going to enlighten the implementation of the proposed CKC-based algorithm and show that, in spite of its simplicity, the obtained recognition rates are comparable with much more sophisticated procedures.

All our experiments comprised the following words: “stop”, “levo” (left), “desno” (right), “naprej” (forward), “nazaj” (reverse), “ena” (one), “dva” (two), “tri” (three), “štiri” (four), “pet” (five); and “šest” (six). The CKC-based recognition was performed using the following parameter values:

- speech segment length $N = 8000$ samples;
- number of artificially introduced observations (by splitting the recorded speech signal), $M = 32$;
- extension factor $M_e = 20$;
- threshold for the similarity of activity indexes, $T_m = 0.07$.

We recorded a population of 20 instances of each of the abovementioned words. The first 10 repetitions were used in the learning phase for a construction of the correlation matrices $\mathbf{R}_{ye}^{(l)}$. The remaining 10 repetitions were included in the recognition process as unknown words.

Fig. 1 depicts the truncated and aligned speech segments of the words “stop” (top), “nazaj” (middle), and “šest” (bottom). These “unknown” segments were compared with all reference words by computing their activity indexes. The results are depicted in Fig. 2. Separate subfigures belong to “stop” (2.a), “nazaj” (2.b), and “šest” (2.c). Every subfigure show 11 plots, for each reference word one. These plots belong to the corresponding activity indexes calculated according to Eq. (10). The index of the correct reference word is exposed by a dotted line in all cases. It is clearly visible that the activity indexes of correct references compute the minimum means for all thee given words.

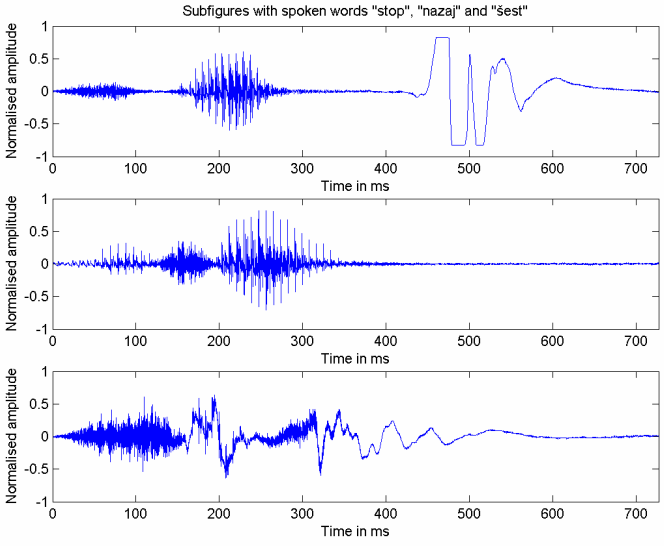

Fig. 1: Time-domain signals of three spoken words: “stop” (top), “nazaj” (middle), “šest” (bottom).

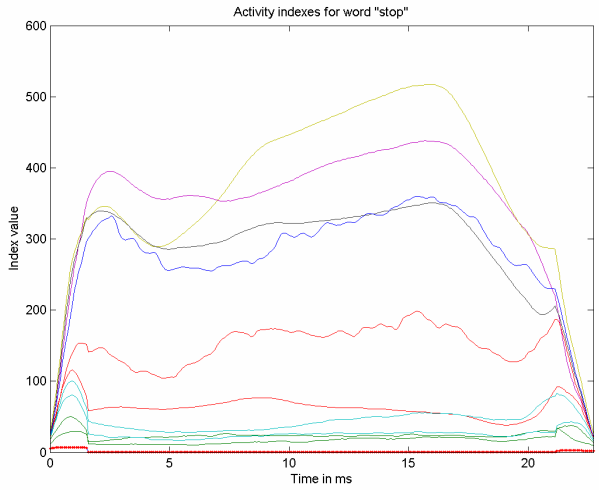

Fig. 2.a: Activity indexes for an example of spoken word "stop"; the lowest dotted (red) line belongs to the correct reference word.

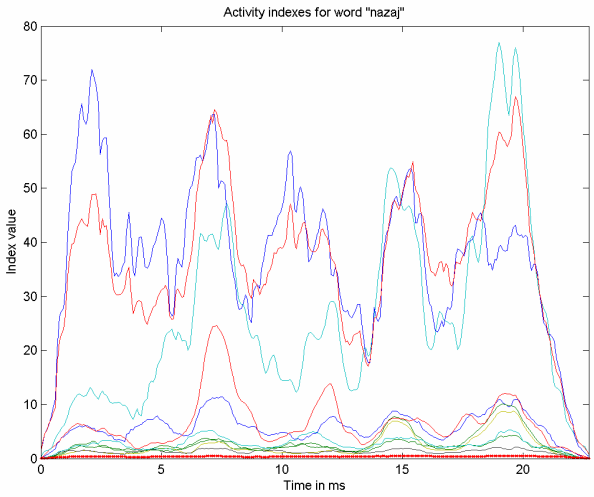

Fig. 2.b: Activity indexes for an example of spoken word "nazaj"; the lowest dotted (red) line belongs to the correct reference word.

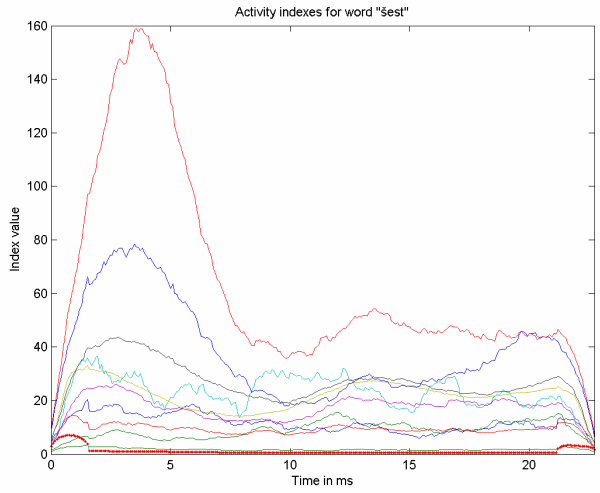

Fig. 2.c: Activity indexes for an example of spoken word "šest"; the lowest dotted (red) line belongs to the correct reference word.

The average recognition rates for the whole vocabulary under investigation are gathered in Table 1, the second column. Table 1 also depicts the recognition results for the DTW-based and NN-based approaches as reported in [5]. It is to be emphasised that the experiments in conjunction with [5] were conducted on larger populations of the 11 words under investigation. The corpus in the learning phase was 10 instances for each word when applying DTW, and from 20 to 80 when applying neural networks. The subsequent recognition was carried out with 60 instances of each class of words.

Table 1: Recognition rates for 11 Slovenian words comparing the recognition rates for the proposed CKC approach with NN- and DTW-based methods; CKC and DTW learnt on 10 instances of each reference word, while NNs used 80 instances. The recognition rates are reported for a corpus of 10 speech segments for each reference in the case of CKC, whereas DTW and NNs worked on 60 segments each.

Spoken words	**Word recognition rate using CKC**	**Word recognition rate using NNs**	**Word recognition rate using DTW**
"stop"	100.00	78.33	93.33
"levo"	100.00	83.33	100.00
"desno"	80.00	81.66	98.33
"naprej"	100.00	25.00	100.00
"nazaj"	70.00	81.66	100.00
"ena"	100.00	76.66	100.00
"dva"	100.00	8.33	93.33
"tri"	90.00	91.66	100.00
"štiri"	100.00	70.00	100.00
"pet"	50.00	78.33	90.00
"šest"	100.00	70.00	100.00
Average recognition rate	**90.00**	**67.12**	**97.72**

5 Discussion and Conclusions

The three speech processing methods compared in this paper rely on a two-phase recognition: firstly, reference templates are built and, secondly, an unknown speech segment is mapped onto the space of those templates. Different distance measures in this space are used to find a minimum-distance template. Both processing phases must take care of the nonstationarity of the speech.

In the case of a limited vocabulary, the reference templates correspond to the classes of the words from this dictionary. Our experiments dealt with 11 Slovenian words selected to control a wheelchair.

DTW- and NN-based recognition commences with MFCC and generates sets of 13 cepstral coefficients (CC). These contain the most important speech information, i.e. the information on the vocal tract response and the pitch, obtained through consecutive speech epochs within the selected frequency bands, which evades the nonstationarities. The recognition phase is different for the two approaches. DTW tries to warp the unknown sets of CC along the corresponding sets of reference templates. The established minimum-length path indicates the closest reference word. On the other hand, NN-based decision implements Kohonen's self-organising maps (SOM) in order to decide the most probable reference word.

Our CKC-based recognition algorithm does not enter the cepstral domain, but tries to align the stationary epochs of a speech segment into a multichannel signal structure. This is then treated by blind system identification based on the MIMO modelling. Firstly, the reference templates are built as the correlation matrices of aligned multiple speech-segment epochs. The recognition, however, is implemented by a new measure based on the CKC approach, explained in the previous sections.

The proposed solution is rather simple, but in the case of a very limited vocabulary the obtained results are quite comparable to those obtained sophisticated recognition methods. DTW does somewhat better, which can be explained by its inherent adaptability to, and an optimal trade-off for, the variability inside and among the speech segments containing the same spoken word.

Although the SOM training was done with 80 instances of each word class, and the correlation matrices for our CKC-based approach were built only with 10 instances, CKC outperforms NNs. The obtained average recognition rate for NNs is at 67.12 %, while for CKC it yields 90.00 % (see Table 1). DTW surpasses other two approaches owing to the reasons already mentioned and achieves up to 97.72 % of recognised words, on average.

Some interesting phenomena have been observed when developing the CKC-based recognition. The activity indexes, for example, have approximately the same value in all the cases when their calculation implies a correct reference. At the same time, they are also very smooth, with a minimum degree of variance. On the other hand, when the correlation matrix, i.e. the reference template, is not the correct one, the indexes tend to increase their average values up to several ten times and become much wavier (see Fig. 2). Further research will be aimed at a deeper understanding of those phenomena.

Acknowledgment

This work was partially supported by the Slovenian Research Programme Funding Scheme P2-0041.

References:

[1] D. Vlaj, Z. Kačič, B. Horvat, "Voice Activity Detection Based on Autocorrelation of the Frequency Spectrum," *Electrotechnical Review*, Vol. 71, No. 4, 2004, pp. 165-170.

[2] T. Fukada, K. Tokuda, T. Kobayashi, S. Imai, " An Adaptive Algorithm for Mel-Cepstral Analysis of Speech," *Proceedings of ICASSP*, 1992, Vol. 1, pp. 137-140.

[3] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi, "Multi-Space Probability Distribution HMM," *IEICE Transactions on Information and Systems*, Vol. E85-D, No. 3, 2002, pp. 455-464.

[4] G. Pačnik, *Prepoznava govora z nevronskimi mrežami (Speech Recognition Using Neural Networks)*–a diploma thesis. Maribor: University of Maribor, 2005.

[5] G. Krebs, *Ugotavljanje uspešnosti metod za prepoznavo govora z omejenim slovarjem (The Speech Recognition Methods Performance Evaluation for a Limited Vocabulary)*–a diploma thesis. Maribor: University of Maribor, 2005.

[6] A. Holobar, D. Zazula, "A Correlation-Based Approach to the Multichannel Blind Decomposition of Binary Sources," submitted to *IEEE Transactions on Signal Processing.*

[7] D. Zazula, "Blind Source Separation Base on a Single Observation," submitted to *EUSIPCO*, Antalya, Turkey, Sept. 2005.

[8] C. S. Myers and L. R. Rabiner, "A comparative study of several dynamic time-warping algorithms for connected word recognition," *The Bell System Technical Journal*, Vol. 60, No. 7, Sept. 1981, pp. 1389-1409.

[9] P. Zeggers, *Speech Recognition Using Neural Networks,* University of Arizona, 1998.