EXAMPLE-BASED REALIZATION OF ISOLATED WORDS RECOGNIZER UNDER LIMITED TRAINING DATA CONSTRAINT

Petr Zelinka[1], Milan Sigmund[1], Detlef Richter[2]

[1]Brno University of Technology, Czech Republic,
[2]Wiesbaden University of Applied Sciences, Germany
xzelin06@stud.feec.vutbr.cz, sigmund@feec.vutbr.cz, richter@informatik.fh-wiesbaden.de

Abstract: Today's speech recognition systems are mostly based on Hidden Markov Models which are known to have good modeling ability based on a training database. Such a database is required to be very large, often many hours of recordings and it must contain all speech units of interest with proper transcriptions. This article describes an isolated words small vocabulary speaker-dependent recognizer which is designed to avoid the need of excessive amount of training data. The recognizer uses example-based approach built around the dynamic time warping technique and k-nearest neighbors classification with additional preprocessing via recurrent artificial neural network and gaussian mixture model-based endpointer.

1 Introduction

The most challenging continuous-speech large-vocabulary recognition systems rely on the generality and fast evaluation of Hidden Markov Models [1], which are indeed the most successful platform for the large-extent recognizers. In the field of small vocabulary isolated words recognition with emphasis on user-dependency can be though found other viable methods [2, 3] that are specific in this area due to their limited extensibility to larger systems. A remarkable quality of these approaches is their tolerance to the limited amount of training data. The training database can moreover be specifically focused only on the needed units, e.g. the set of words to be recognized by the system. Such a limited database is easily obtainable from the target user(s) of the system for the specific set of word commands.

The basic principle of limited-data word recognition methods is the direct comparison of the received sample against the available database. The natural inequality of repeated realizations of the same word in the sense of variable duration and uneven lengths of different intra-word sections calls for a robust time-alignment techniques, among which the most popular is the dynamic time warping (DTW) [4-6]. This algorithm allows comparison of two words in the segment-wise manner using any of the known distance-measurement approaches. If the global distance to one of the word classes falls below a given threshold, it is identified as the specific word.

Reliable operation of this kind of classifier requires that the input utterances are precisely scaled in amplitude to fit the given stored realizations as closely as possible. Another issue to deal with is the background noise, which is added to the speaker's signal. This noise must be taken into accounted and, if possible, reduced without intolerable distortion of the resulting data [7, 8].

Following chapters describe individual parts of a spoken commands recognizer, which is currently being developed in our lab and preliminary evaluations of some performance characteristics.

2 Preprocessing of input utterances

2.1 Signal parameterization, noise reduction

Input speech is digitized using sampling frequency of 16 kHz, which is considered an optimal compromise between ability to capture high-frequency phonemes (fricatives) and need to store too extensive amounts of data [9]. The stream of sound samples is then segmented in 1024-sample chunks with 75% overlap. This allows precise tracking of fast-varying spectral changes whilst giving good spectral resolution. Individual segments are parametrically described using feature vectors composed of the mel-frequency cepstral coefficients (MFCC) [10, 11]. Prior to construction of MFCCs, spectral representation of the input signal is treated for additive noise using the quantile-based spectral subtraction (SS) [8].

SS estimates power spectral density (PSD) of background noise from silent segments between the words. This estimate is then used to eliminate detected noise levels in all spectral bins throughout the speech. Application of this operation can be described as

$$\hat{S}(k,t) = \max\{X(k,t) - \alpha\hat{N}(k,t), \beta X(k,t)\}, \quad (1)$$

where $X(k,t)$ is the k-th PSD of the noise-corrupted input signal in frame t, $N(k,t)$ represents the k-th estimated noise PSD and $\hat{S}(k,t)$ is the PSD of the denoised signal. Constants α and β control the amount of noise being removed from the signal.

2.2 Voice signal strength normalization, dynamics compression

Conventionally, speech signal magnitude is normalized either in time domain by an automatic gain control (AGC) algorithm or in parametric domain using various normalization techniques [7].

We propose a method of level control based on well-known masking effects of human hearing [12]. The method operates in log-mel-filter-frequency domain (i.e. the last step of MFCC calculation process just before the discrete cosine transform is applied), since this signal representation is the closest approximation of human way of sound interpretation in early stages of neural processing [12]. The basic idea is to track local maxima and minima of mean coefficient values across all frequencies for each segment and enforce that all values fall within a pre-arranged interval, which was chosen to be (0; 1). This span was motivated by the need for the feature vector stream to be usable by the subsequent neural network processing both as an input and as a teacher-forced output during the learning (see below).

The dynamics compression of the log-mel-spectral coefficient vector stream is divided into two stages: firstly, a pair of attack/release look-ahead dynamics compressors is employed separately based on mean maxima and mean minima delineating the upper and lower envelope contours. These contours determine the shift and stretching of all feature vectors' values to occupy the symmetrical (-2; 2) interval, which is then soft-clipped using a sigmoidal function to the final (0; 1) range. In order to be able to softly cut-off extensive peaks, the upper contour uses two serially chained compressors – the first one operates in a slower fashion defining overall dynamics and the second one is used only to iron out excessive peaks above the +2 boundary (application of the second stage assumes the shift and stretch have already been carried out). Both minimal and maximal contour boundaries are subject to additional constraints based on observed signal properties to avoid excessive amplifying of background noise or picking of too weak random sounds. Lower contour boundary is deliberately set a bit higher than would correspond to the measured value to induce additional mild background noise suppression. Figures 1 and 2 illustrate operation of described

algorithms on a signal sample comprising of three repetitions of the same word having large differences in amplitude for each of the three realizations.

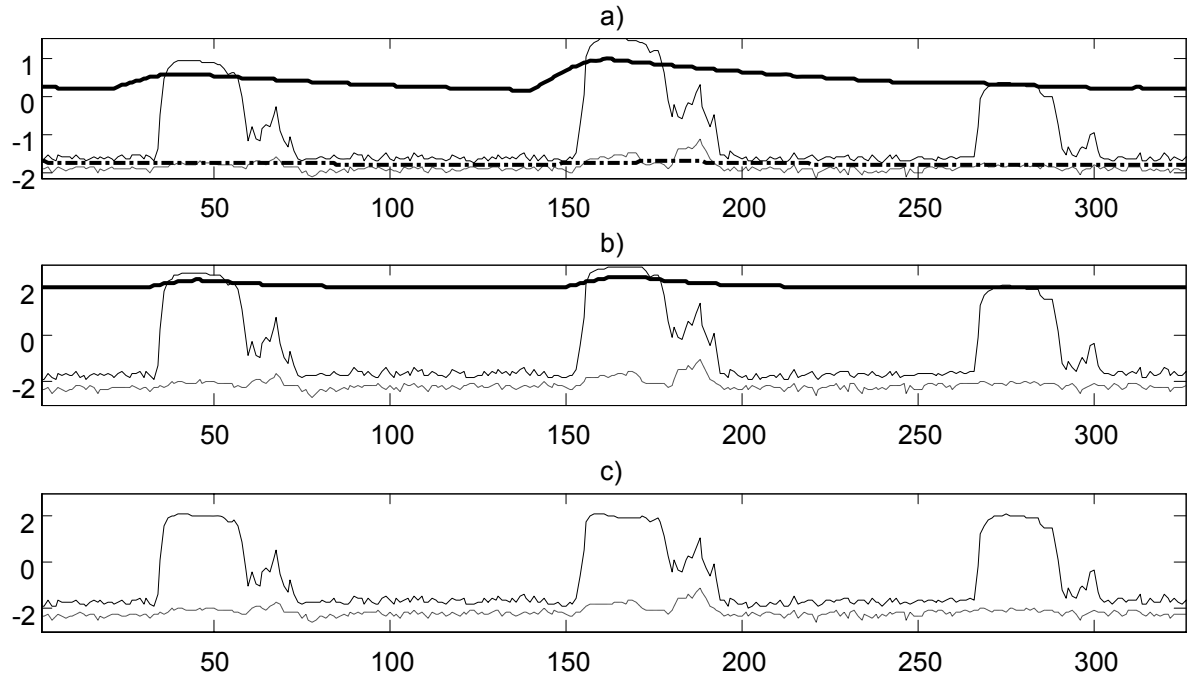

Figure 1 - Operation of the dynamics compressor in logarithm-mel-frequency domain. Thin solid line shows the mean feature vectors' maxima, thin dash-and-dot are the minima. The thick lines (solid and the dash-and-dot) correspond to the upper and lower envelope contours as are found by the dynamics compressors. Graph a) shows the situation on the input of the compressor, in graph b) have all the values already been shifted to the (-2; 2) interval and finally in graph c) all excessive peaks have been ironed out by the second stage compressor. Note that the lower contour boundary in a) is slightly above the background noise level to avoid amplification of noise.

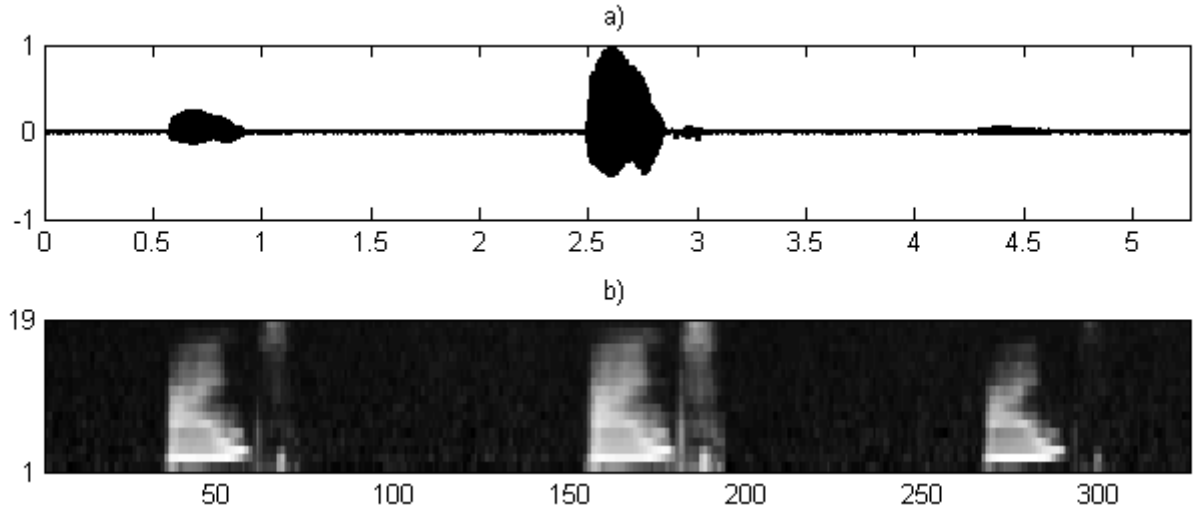

Figure 2 - a) Input signal in time domain (time axis in seconds), b) feature vectors after the dynamics compression with x-axis in segments and y-axis corresponding to the mel-frequency.

2.3 Speech conversion

Optimal recognition accuracy requires that the database speech samples are as close to those received from the speaker as possible. It is unfeasible to include every speaker to the DTW-based classifier due to sharply increasing computational complexity as the number of database samples rises. Therefore it is more advantageous to alter just received speech to bring it closer to the available word samples. Since the system uses one reference speaker, all the other speakers' utterances need to be modified to match the reference speaker's characteristics. This is often referred to as speaker adaptation or speaker normalization. Often used techniques are vocal tract length normalization [13], cepstral mean and variance normalization [14], various kinds of spectral transformations [15, 16] and finally utilization of artificial neural networks (NN) with various strategies exploiting great potential of these generalized structures.

Several authors used NNs trained on pairs of samples with one belonging to the reference speaker and the other one to the speech to be converted [17, 18]. Often used multilayer perceptron structure, however, neglect the fundamental property of human speech being a continuous stream of timely-interdependent elements which cannot be simply chopped into autonomous units. Hence our solution is to employ fully interconnected recurrent neural network (RNN) [18] with 150 nodes. This kind of NN has the ability of exploiting both spatial and temporal patterns (theoretically to unlimited time depth) in order to produce the desired output pattern for current time frame.

Teacher-supplied outputs are obtained by DTW aligning of two equal speeches given by a speaker to be learned and by the reference speaker. RNN therefore learns to convert parameterized speech of one speaker to "look like" being produced by the other one. The best-matching pairs of the two speeches found by DTW are freed of duplicates; hence not all segments in the input speech for the RNN have assigned prescribed outputs. This gives the RNN additional challenge of treating unmarked segments in a right way leading to global path following marked segments. Additionally, each prescribed output is delayed by a fixed amount of segments to give RNN necessary time to observe certain amount of incoming speech stream before it is required to decide about what output should be produced. This allows it to overcome glitches or imperfectly uttered phonemes based on the following phoneme continuation. Learning algorithm used is the backpropagation through time (BPTT) with weight update procedure according to that proposed in [19] giving the fastest convergence for large networks. The teacher-forcing technique is used to further speed up the training process. Each speaker provides two continuous utterances: one as the data on which the RNN is trained and one as the verification testing data to prevent over-fitting. The training procedure is carried out until a minimum overall error for the testing utterance is reached assuring maximal generalization performance of the resulting RNN. This over-fitting avoidance is crucial due to the fact, that the training data is composed only of short amount of speech (typically a few minutes) in agreement with overall philosophy of proposed system.

Figure 3 shows feature vector sequences corresponding to 5 spoken words illustrating conversion of a given speaker's speech by the RNN to bring it closer to the reference speaker's utterance of the same words. An interesting side-effect of RNN's operation is robust removal of various random non-speech sounds (breathing, pops/clicks etc.), because the input learning data contained plenty of these imperfections but the exemplary reference speaker's speech was hand-edited to enforce silence in pauses between the words. It seems that the RNN learned specific time-space patterns of the speaker's pronunciation to cancel out everything uncommon to his/her articulation.

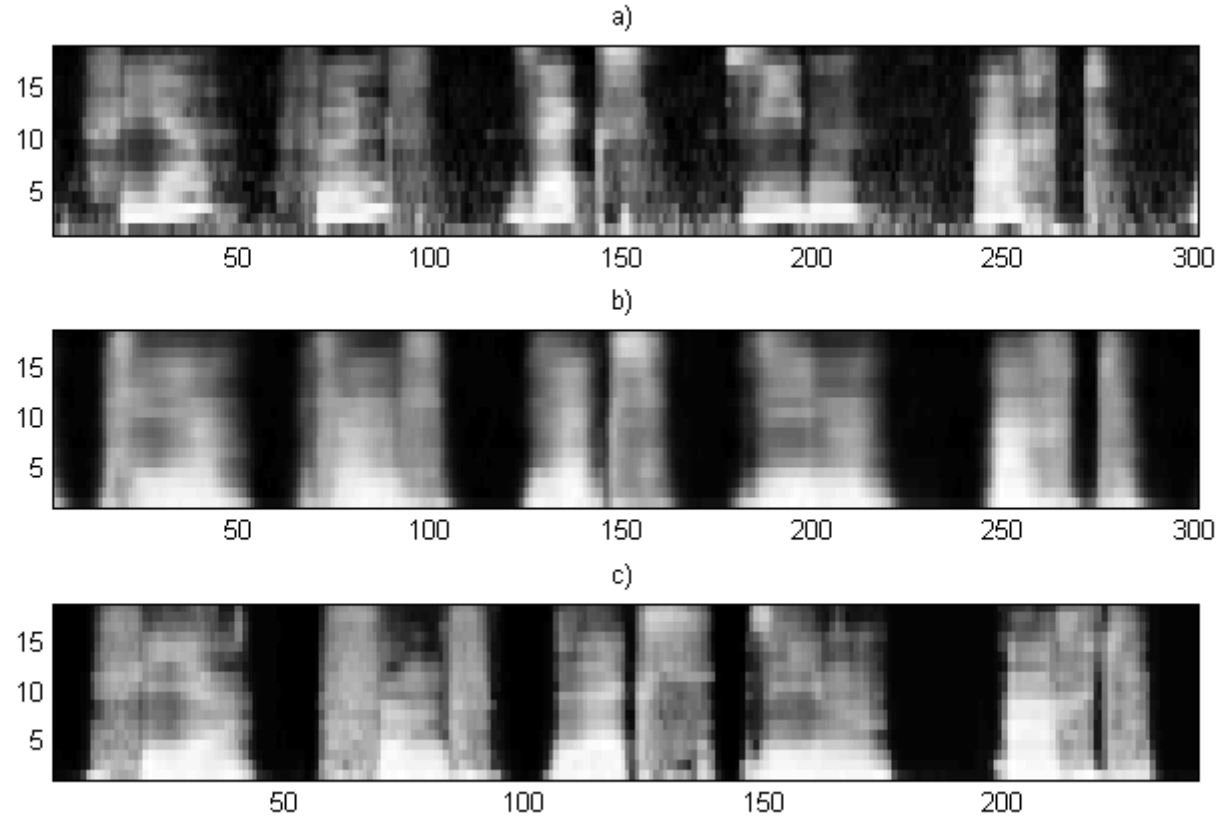

Figure 3 – Parameterized speech of a speaker on the input of the RNN a), output of the RNN b) and the same utterance from the reference speaker c). The RNN is trained to convert a) to look as much as c) as possible in a minimum mean square error sense (speeches are not time-aligned).

3 Detection of word boundaries

Since the DTW-based isolated words recognizer operates on the single-word basis, it is necessary to provide a word-identifying classifier, the endpointer [20]. It was chosen to use a gaussian mixture model (GMM) [18] – based statistical approach with two specific gaussian classifiers working in serial. The first classifier uses two classes: speech vs. non-speech. Both classes have assigned 30-mixture GMMs, whose feature space is composed of MFCCs, their 1st and 2nd derivatives obtained from 3rd order polynomial regression and mean absolute error of the regression fit over neighboring 5 segments. The second classifier is specifically designed to discriminate between the starts and ends of words; thereby it is trained only on boundary points and its two classes are start/end. These classes also uses 30-mixture GMMs, but the feature vector is composed of the 3rd order polynomial coefficients and mean probability density difference before and after the given segment obtained from the previous classifier's outputs in span of 5 segments. This cascade, together with certain time constraints, is able to mitigate many false-positive detections which could occur if only one-stage classification was used. All probability densities are smoothed out by moving average thus random glitches and local miss-classifications have little impact on overall performance. In order to be able to set the balance between false positives and misses, the sequence between a given start-end pair must have higher-than-threshold average value of the log-likelihood of speech minus silence class.

We are currently investigating means of increasing robustness of the endpointer by constructing a set of GMMs for a range of signal-to-noise (SNR) values. Since it is

impractical to have too many such models, it was decided to store 6 settings equally distributed over the supported SNR range. The actual probability density value for arbitrary SNR of this range is then calculated by Hermite spline interpolation of these PDFs in a point given by the current feature vector. Figure 4 shows measured results of the word-spotting accuracy for a range of additive white gaussian noise (AWGN) SNRs (most of them falling between the trained values thus relying on the described interpolation of PDFs). No false positives were detected; amount of misses is zero above SNR = 4.

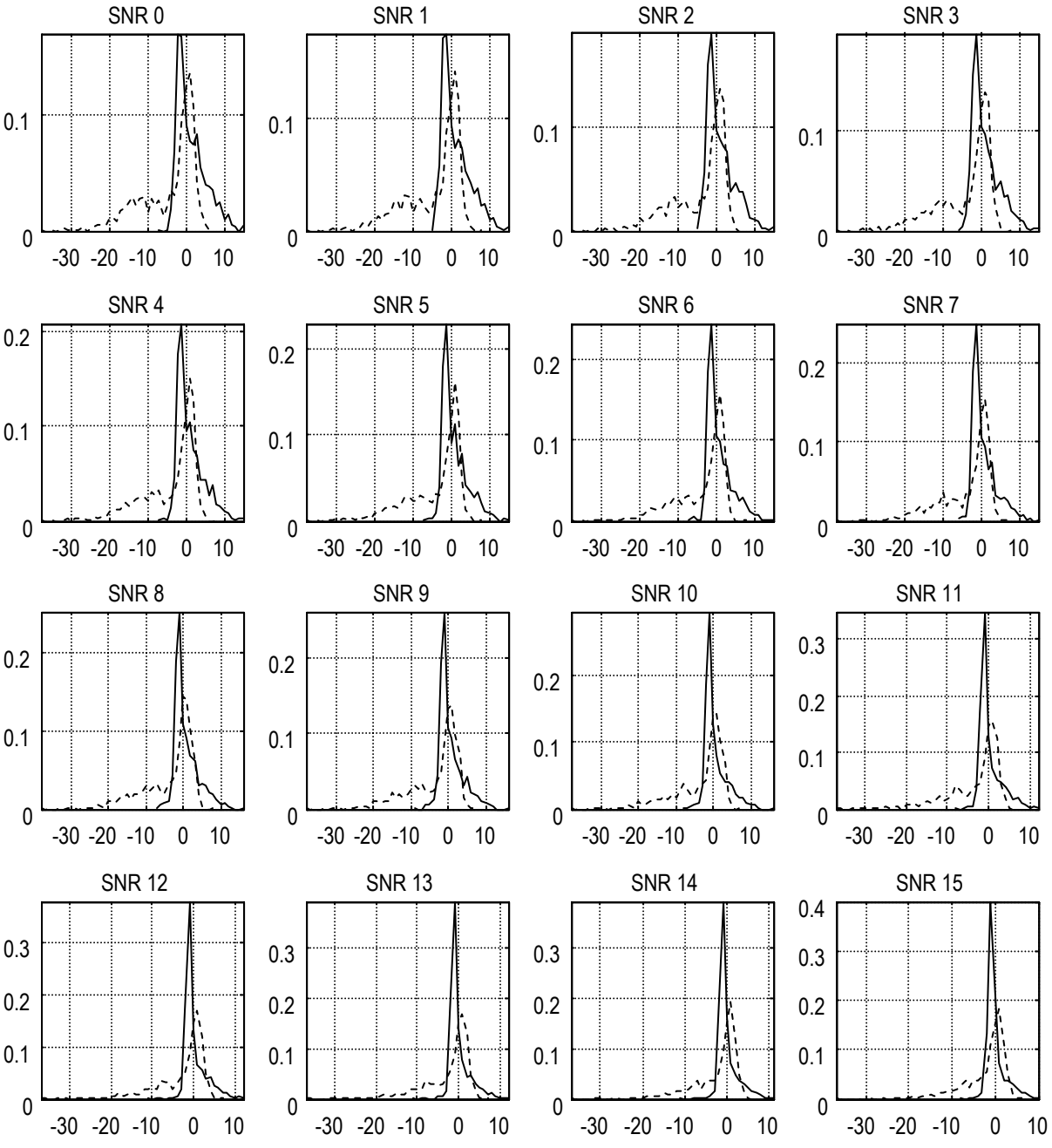

Figure 4 – Histograms of startpoint (solid line) and endpoint (dashed line) variations in units of segments (roughly 16 ms per segment) for a range of testing SNR values of AWGN.

Obviously, AWGN poses little challenge to the classifier; further work will therefore concentrate on more complicated types of noise (car, bubble etc.). Extension to other noise types could be achieved by multi-dimensional interpolation in a similar fashion to the current design. A precise noise type identifier will play a significant role.

This classifier clearly depends on the amount of training data, which could suggest a contradiction to the whole system's philosophy. The simplest two-class case, however, allows decent characteristics even when trained only on relatively few examples given by small group of speakers.

4 Word classification

As already mentioned, classification of received spoken word commands is carried out using dynamic time warping of stored words towards received utterances and measuring the global distance of these pairs. Among these distances the most probable word class is identified via the soft-voting k-nearest neighbors method [18], forming the core of the example-based approach. DTW uses intersection of Sakoe-Chiba band and Itakura parallelogram as the global constraint; the dynamic programming path is also subject to symmetrical local constraint defining maximal allowed number of steps in one of two viable off-diagonal directions. For measuring similarity of aligned segments the Mahalanobis distance measure [2] is utilized. A unique covariance matrix for each segment of every stored word is employed to cover natural variances of different phonemes within a word. These matrices are found during construction of the classifier prior to its usage by comparing each example word against all the other ones available from the same class. This slight detour from strictly example-based approach is awarded by better recognition accuracy compared to any fixed measures (absolute distance, Euclidian distance etc.). The whole classifier module is still in early stages of research, therefore no exact representative results are available yet. Current work is focused on improving discriminability of competing classes by utilizing more of the time-related information in speech by varying local DTW constraints in accordance with observed dispersion of word sub-units. This, however, might lead to increased demands on the amount of available training data eliminating the primary advantage of the system.

5 Conclusion

A basic structure of an isolated-words small vocabulary speaker-dependent isolated words recognizer was outlined. The design philosophy is aimed at achieving decent recognition accuracy without the need of excessive amounts of training data. Several key points necessary to overcome performance drops caused by speech variations were described: a dynamics compressor for ensuring consistent feature vector values, recurrent neural network for alleviating speaker-dependent differences and noise-robust endpointer for identyfying word boundaries. The whole system is yet in early stages of development and is expected to grow in complexity as the need for higher reliability will identify concrete weak points.

Acknowledgements

Research described in the paper was financially supported by the Czech Grant Agency under grant No. 102/08/H027 and by the research program MSM 0021630513 Advanced Electronic Communication Systems and Technologies (ELCOM).

References

[1] Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. In: Proceedings of IEEE, Vol. 77, No. 2. 1989, pp. 257 – 286.
[2] Rabiner, L., Juang, B.: Fundamentals of speech recognition. New Jersey: Prentice Hall 1993.
[3] O'Shaugnessy, D.: Speech Communication. Indianapolis: Addison-Wesley Publishing Company 1987.
[4] Myers, C., Rabiner, L., Rosenberg, A.: Performance tradeoffs in dynamic time warping

algorithms for isolated word recognition. In: IEEE Transactions on ASSP, Vol. 28, pp. 623 – 635.
[5] Itakura, F.: Minimum prediction residual principle applied to speech recognition. In: IEEE Transactions on ASSP, Vol. 23, pp. 52 – 72.
[6] Rabiner, L., Rosenberg, A., Levinson, S.: Considerations in dynamic time warping algorithms for discrete word recognition. In: IEEE Transactions on ASSP, Vol. 26, pp. 575 – 582.
[7] Wchuller, B., Wollmer, M., Moosmayr, T., Rigoll, G.: Recognition of noisy speech: a comparative survey of robust model architectures and feature enhancement. In: EURASIP Journal on Audio, Speech and Music Processing. 2009, pp. 1 – 16.
[8] Stahl, V., Fischer, A., Bippus, R.: Quantile based estimation for spectral subtraction and wiener filtering. In: Proceedings of ICASSP. 2000, pp. 1976 – 1978.
[9] Sanderson, C., Paliwal, K.: Effect of different sampling rates and feature vector sizes on speech recognition performance. In: Proceedings of TENCON 97. 1997, pp. 161 – 164.
[10] Mermelstein, P.: Distance measures for speech recognition, psychological and instrumental. In: Pattern recognition and artificial intelligence. New York: Academic, pp. 374 – 388.
[11] Davis, S., Mermelstein, P.: Comparsion of parametric representations for monosyllabic word recognition in continuously spoken sentences. In: IEEE Transactions on ACSSP, Vol. 28, pp. 357 – 366.
[12] Divenyi, P.: Speech separation by humans and machines. Boston: Springer + Business Media, Inc. 2005.
[13] Metze, F.: Discriminative speaker adaptation using articulatory features. In: Speech Communication, Vol. 49, Issue 5. Amsterdam: Elsevier 2007, pp. 348 – 360.
[14] Viikki, O., Laurila, K.: Cepstral domain segmental feature vector normalization for noise robust speech recognition. In: Speech Communication, Vol. 25, Issue 1-3. Amsterdam: Elsevier 1998, pp. 133 – 147.
[15] Choi, H., King, R.: Speaker adaptation through spectral transformation for HMM based speech recognition. In: Proceedings of SIPNN, Vol. 2. 1994, pp. 686 – 689.
[16] Bellegarda, J., Souza, P., Nadas, A., Nahamoo, D., Picheny, M., Bahl, L.: The metamorphic algorithm: a speaker mapping approach to data augmentation. In: IEEE Transactions on speech and audio processing, Vol. 2, Issue 3. 1994, pp. 413 – 420.
[17] Huang, X.: Speaker normalization for speech recognition. In: Proceedings of the ICASSP 92, Vol. 1. 1992, pp. 465 – 468.
[18] Katagiri, S.: Handbook of neural networks for speech processing. Boston: Artech House 2000.
[19] Robinson, A.: An application of recurrent nets to phone probability estimation. In: IEEE Transactions on neural networks. Vol. 5, No. 2. 1994. pp. 298 – 305.
[20] Yamamoto, K., Jabloun, F., Reingard, K., Kawamura, A.: Robust endpoint detection for speech recognition based on discriminative feature extraction. In: IEEE Transactions on ASSP, Vol. 1. 2006, pp. 14 – 19.