Impact of choice of training data and patterns reduction in speaker dependent speech recognition

Jakub Gałka, Tomasz Jadczyk, Bartosz Ziółko, Dawid Skurzok

Department of Electronics
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
{jgalka,jadczyk,bziolko,skurzok}@agh.edu.pl

Abstract

The paper presents an optimisation process of training data content for a speaker dependent ASR task. Each set of patterns representing speech should cover entire space of possible observations, but on the other hand it should not contain unnecessary data because of time efficiency and possible unbalance of the model if any part of the space is overrepresented. Chosen feature space reduction methods were described as well.

Keywords: speech recognition, speech feature space reduction

1. Introduction

A word decoder is a very crucial part of each automatic speech recognition system (Young et al., 2005; Ziółko et al., 2011). It has crucial impact not only on quality of final recognition but also on time efficiency because this function is called several times in continuous speech recognition task. Our speech recognition system is based on perceptual wavelet decomposition (Ziółko et al., 2010b).

Dynamic Time Warping (DTW) (Rabiner and Juang, 1993) and Modified Weighted Levenshtein Distance (MWLD) (Ziółko et al., 2010a) algorithms were tested in several different training scenarios for one particular testing speaker, using his recordings and recordings of other speakers as well. Several different scenarios were addressed - with removing some of the patterns from the model and with recognition using all of them. The third scenario was the model trained only using data of the testing speaker. In some cases only 5 CORPORA speakers were used, in others all 26 male speakers. Similar experiments were conducted for the test speaker "small" (around 20 minutes) and "large" (around 40 minutes). Several approaches to removing patterns form the database are compared as well.

Over the past decades there has been considerable research effort devoted to reduction of features space. Nonlinear discriminant analysis was successfully applied (Hu and Zahorian, 2010). It provides reduction by introducing a neural network and Principal Components Analysis (PCA) which process data before applying the hidden Markov models. PCA performs a Karhunen-Loeve transform in order to reduce the correlation of the network outputs and improve their suitability for gaussian mixture models.

Features dimension based on a quantitative model of the human auditory periphery was described (Koniaris et al., 2010). The method maximises similarities of geometries of subset feature and a spectro-temporal auditory model capturing frequency and time domain masking.

This work was supported by MNiSW resources for science as statutory activity.

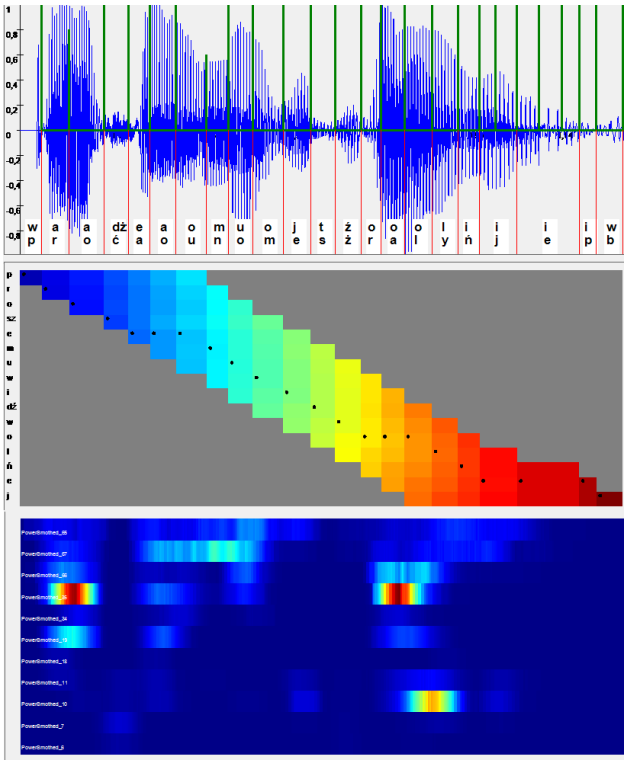

Figure 1: Printscreen of our wavelet based speech recognition system (Ziółko et al., 2011), tested by the experiment described in the paper

A random projection can be also used for dimensionality reduction (Takiguchi et al., 2010), where the original set of data is projected onto a subspace by a random matrix. It preserves approximate values of Euclidean distances between points in the space.

The next section presents the automatic evaluation of usefulness of phoneme patterns we suggested. The corpora we used are described in section 3. Section 4 depicts results of our experiments. The paper is summed up with conclusions.

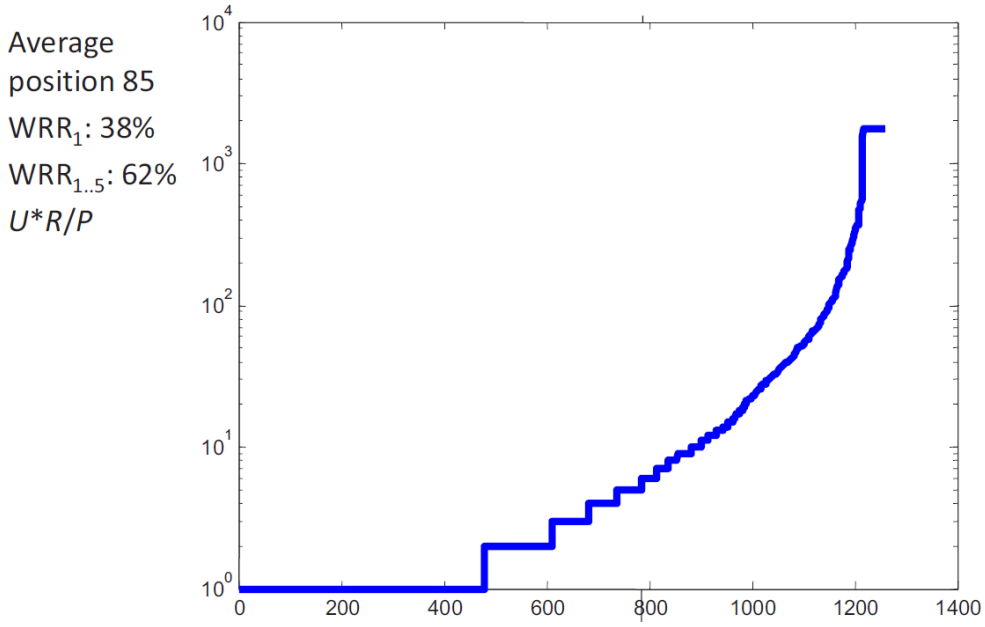

Figure 2: Distribution of a correct word in a hypotheses ranking (horizontally number of a test recording, vertically position of the word in ranking) in the DWT experiment

2. Speaker adaptation by unsupervised learning of patterns usability

The method conducts speech recognition and evaluates particular patterns giving rewards R (for participating in correct recognition) and penalties P (for participating in wrong hypotheses which were given as the most likely). The number of times the phoneme pattern was used is also noted (as U). These three parameters allows to evaluate each pattern. However, it is not obvious how to do it. Four different formulae for patterns evaluation were applied

$$E_1 = (R - P)U \ , \tag{1}$$

$$E_2 = R - P \ , \tag{2}$$

$$E_3 = U\frac{R}{P} \ , \tag{3}$$

$$E_4 = \frac{R}{P} \ . \tag{4}$$

Results of impact of using patterns' space reduction by applying (1) - (4) on recognition are presented later on in the paper.

3. Data

The main part of training data was CORPORA (Grocholewski, 1995). Its speech files were recorded with the sampling frequency $f_0 = 16$ kHz equivalent to sampling period $t_0 = 62.5$ μs. Speech was recorded in an office with a working computer in the background, which makes the corpus not perfectly clean. Signal to noise ratio is not stated in the description of the corpus. The database contains 365 utterances:

- 33 single letters,
- 10 digits,
- 200 names,
- 8 simple computer commands,
- 114 short sentences.

Each of these were spoken by 11 females, 28 males, and 6 children (45 people), giving 16,425 utterances in total. One set spoken by a male and one by a female were hand-segmented. The rest were segmented by the dynamic programming algorithm, which was trained on hand-segmented ones, and manually checked afterwards. The quality of all transcriptions can be assumed to be as good as hand-made transcription.

Experiments with a male speaker were conducted, so only male CORPORA speakers were used in trainings. In some scenarios we used all 26 speakers, while in some others only 5 picked randomly. Regarding the speaker data, we had around an hour of his recordings, manually transcribed into words. They were transcribed using forced alignment into phonemes. In cases with index [1] in Tab. 1 only 5 CORPORA speakers were used as a support for forced alignment and in case of [2] 26 speakers were used. Different microphones and different word content were used comparing to CORPORA. Part of this data was used for training and part for testing.

4. Experiments and results

An automatic speech recognition experiment was conducted on 500 word recordings using AGH ASR system (Ziółko et al., 2011). The test recordings were different then the training ones. Results are presented for various settings of parameters and corpora used for training.

Tests using MWLD (Ziółko et al., 2010a) were conducted as well. The best results were obtained for the optimal choice of the decoder:

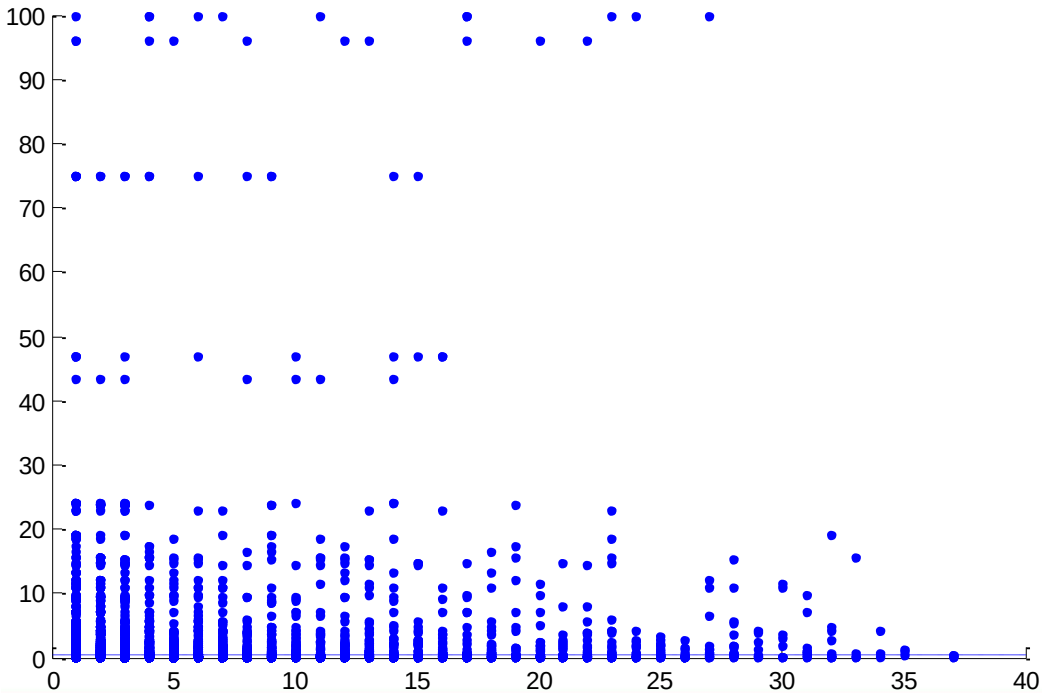

Figure 3: Image of statistics of phoneme hypotheses for the correct phonemes (vertically) against position of the correct word expressed as % of a position in a dictionary (horizontally), where 100 is the worst. The figure shows that good phoneme hypotheses are corelated with good word hypotheses

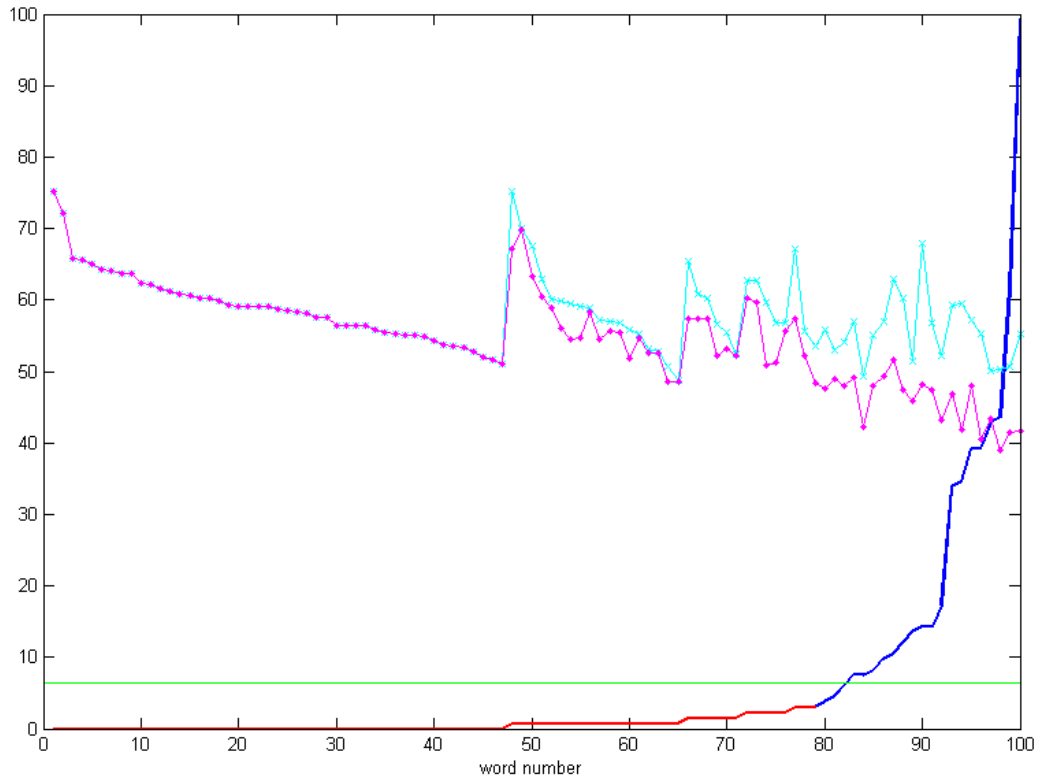

Figure 4: Results for MWLD. Vertically % of test recordings. The bold red-blue line shows distribution of position of the correct word in the ranking of hypotheses (Red in top 5, blue - the rest). The cyan line with points represents a probability of the correct word, while the magenta one with crosses shows the probability of the strongest hypothesis.

Table 1: Comparision of results in word recognition rate for different training data and a criterion of removing patterns from the database. Models were tested on 500 word recordings with 2270 words in a dictionary with 2914 phonetic transcriptions. R stands for a reward, and P stands for a penalty. In both cases they are accumulated values of likelihoods from hypotheses were particular phoneme pattern helped to recognise correctly (R) and introduced error (P). U stands for the number of times the phoneme pattern was used. *Small* is a smaller speaker adaptation set of recordings and *large* is a full one. [1] stands for transcriptions during forced alignment only 5 CORPORA speakers were used and [2] stands for 26 speakers used for this aim. Table presents % of recordings with the correct word in a list of 5-best hypotheses

Training data	-	(R-P)*U	R-P	U*R/P	R/P
Only CORPORA (5 speakers)	48.60%				
CORPORA(5 speakers)+ small	66.80%	46.00%	50.80%	60.00%	37.20%
Only small[1]	66.80%	64.60%	64.60%	64.60%	64.60%
CORPORA(5 speakers)+ large	70.00%	33.20%	35.00%	52.00%	20.60%
Only large[1]	65.40%	64.80%	64.60%	66.60%	64.80%
Only CORPORA (26 speakers)	57.60%				
CORPORA(26 speakers)+ small	68.20%	60.00%	60.20%	60.20%	60.20%
Only small[2]	63.40%	63.40%	63.40%	63.40%	63.40%
CORPORA(26 speakers)+ large	70.00%	57.60%	56.60%	62.40%	54.80%
Only large[2]	63.40%	65.20%	65.60%	63.60%	63.20%

- $P_{del} = 0.3$,
- $P_{ins} = 0.15$,
- $P_{repl} = 0.085$,
- substitution coefficient = 5,
- reduction of substitution cost = 4,
- reduction of deletion cost = 0,
- segmentation into 40 milliseconds.

The presented results (Fig.4) are for CORPORA 5 speakers plus the speaker training data. 79% were words were in top 5 with 49% on the highest position. The test was conducted on a dictionary with 1758 words. The average position of the correct word is 8.

Tests using DTW showed similar behavior. They were made with many more scenarios to compare different strategies of adapting to a speaker (Tab.1). Applying more adaptation data (speaker large) was definitely more efficient. Using the model with 26 other speakers was similarly efficient as with the only 5 speakers model, which would suggest that the speaker data is the most important. But it is not true because the model without CORPORA was much worse. Also the tests showed that in this scenario, removing patterns which received negative grades during adaptation, in all checked ways, decreased recognition.

5. Conclusions

Speaker adaptation by unsupervised learning of patterns usability was presented with experimental evaluation of choice of training data in speaker dependent ASR. The results suggest that the reduction without losing some efficiency of the model to predict content of audio recordings is not possible in the current method, however, the results depend also on amount of data used. The reduction could be more successful if a larger patterns' space was available. The impact of a rate of recordings of the number of other speakers and the speaker being recognised was analysed and described.

6. References

Grocholewski, S., 1995. Założenia akustycznej bazy danych dla języka polskiego na nośniku cd rom (Eng. Assumptions of acoustic database for Polish language). *Mat. I KK: Głosowa komunikacja człowiek-komputer, Wrocław*:177–180.

Hu, H. and S. A. Zahorian, 2010. Dimensionality reduction methods for hmm phonetic recognition. *Proceedings of ICASSP*:4854–4857.

Koniaris, Ch., S. Chatterjee, and W. Bastiaan Kleijn, 2010. Selecting static and dynamic features using an advanced auditory model for speech recognition. *Proceedings of ICASSP*:4342–4345.

Rabiner, L. and B. H. Juang, 1993. *Fundamentals of speech recognition*. New Jersey: PTR Prentice-Hall, Inc.

Takiguchi, T., J. Bilmes, M. Yoshii, and Y. Ariki, 2010. Evaluation of random-projection-based feature combination on speech recognition. *Proceedings of ICASSP*, 2150-2153.

Young, S., G. Evermann, M. Gales, Th. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, 2005. *HTK Book*. UK: Cambridge University Engineering Department.

Ziółko, B., J. Gałka, D. Skurzok, and T. Jadczyk, 2010a. Modified weighted Levenshtein distance in automatic speech recognition. *Proceedings of XVI KKZMBM*:116–120.

Ziółko, M., J. Gałka, B. Ziółko, and T. Drwięga, 2010b. Perceptual wavelet decomposition for speech segmentation. *Proceedings of the INTERSPEECH, Makuhari*:2234–2237.

Ziółko, M., J. Gałka, B. Ziółko, T. Jadczyk, D. Skurzok, and M. Mąsior, 2011. Automatic speech recognition system dedicated for Polish. *Proceedings of Interspeech, Florence*.