ADAPTATION OF UNKNOWN DATA TO ALREADY TRAINED SPEECH RECOGNITION SYSTEM

František GRÉZL, Doctoral Degree Programme (5)
Dept. of Computer Graphics and Multimedia, FIT, BUT
E-mail: grezl@fit.vutbr.cz

Supervised by: Dr. Jan Černocký

ABSTRACT

This work deals with different speech rate for phoneme recognition based on a Tandem system with TRAP-feature processing. The estimation of rate is based on measuring entropy at the output of the phoneme classifier. Preliminary results obtained on TIMIT database show that this technique may lead to estimation of correct speech rate for mismatched data. This work brings further insight on temporal trajectories and systems based on them.

1 INTRODUCTION

In Automatic speech recognition (ASR) we would like to use the same features when dealing with new database or unknown data. This is very easy for standard features based on Fourier spectrum. But when the feature extraction contain parts trained on data, the new database with different characteristic may cause extraction of unreliable features and degradation of ASR performance.

Data trained classifiers are inherent part of TempoRAl Pattern (TRAP) feature extraction [1]. These features are less sensitive to changes of frequency characteristic due the independent processing of frequency sub-bands and their final combination. But they may be sensitive to time variation, such as speech rate (average phoneme duration), since they are using temporal trajectory of logarithmic energy as input.

Training of feature extraction may be very time and resources expensive, so use of already trained feature extraction is preferred. The goal of our work is to eliminate the effect of different speech rate between the training and testing databases which decreases the performance of TRAP system. We hypothesize, that different speech rate has stretching effect on critical band energy trajectory and can be diminished by re-sampling of critical bands spectrogram (CRBS).

The re-sampling can be estimated using average entropy over small set of “unknown” speech data. We assume that the average entropy will give us some kind of measure, how close test data are to data presented in training.

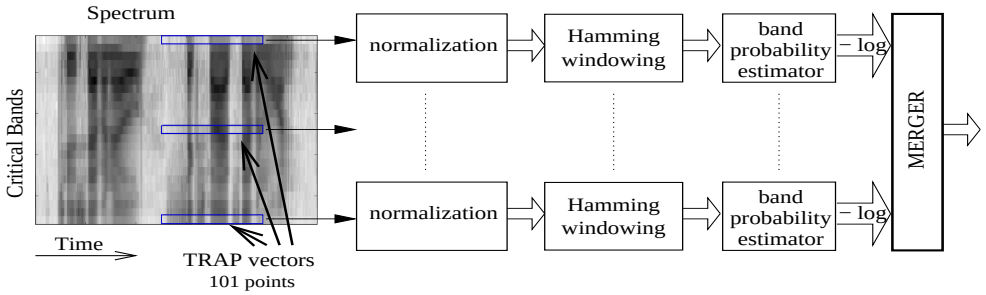

Figure 1: TRAP system

2 TRAP SYSTEM AND RECOGNIZER SETUP

After speech segmentation into 25 ms frames and computing of the power spectrum, spectrum energies are integrated into filter bands (M=15 Bark scaled trapezoidal filters) and logarithm is taken. In each band, actual frame with +/− 50 frames context is taken, so we have 101 points long TRAP vector. Mean normalization of TRAP vectors follows. The Hamming window is applied on the TRAP vector.

The vector at the end of this processing is put into the **band probability estimator** — a three layer neural net. This net is trained to classify the input vector into one of the N classes. The input layer size is equal to the size of input vector (101), one hidden layer with 100 units and the output layer, the size of which is equal to number of classes N (we used 45 phoneme classes). All output vectors are concatenated into a vector M×N points long. This vector goes through negative logarithmic nonlinearity and then forms the input for the **merger probability estimator**. Merger probability estimator is also a three layer neural net trained to classify input vector into the classes — the same target classes as the band probability estimators. The first layer has M×N points, hidden layer has 300 units and the third layer has again N points. Its function is to merge particular band estimations into one final posterior probability vector. The scheme of the TRAP system is shown in Fig. 1.

Negative logarithm is taken and decorrelation using PCA is done on output of the merger probability estimator. This vector creates an input vector for standard HTK based GMM-HMM recognizer which task is to recognize phonemes. The same 45 phonemes are used with no language model. Each phoneme is modeled by 5 emitting states, 32 mixture components per state. The word insertion penalty (wip) was tuned to obtain similar number of insertions and deletions.

3 INITIAL EXPERIMENTS

Initial experiments are done on TIMIT database [2], where the different speech rate was simulated by CRBS re-sampling. This re-sampling is done by skipping (down-sampling) or inserting (up-sampling) frames in CRBS. When frame is inserted, its values are linearly interpolated form previous and following frame.

First the train part (4620 sentences) CRBSs were up-sampled with coefficient 1.65. The TRAP feature estimators and HMM recognizer (further referred as *system*) were trained on this data . Then the test data CRBSs (1680 sentences) were up-sampled with the same

coefficient and given to the system. This experiment gives the performance for `matching data` (train and test) . The phoneme recognition accuracy is shown on first line in Tab. 2.

In the next experiment, the original data were given into the system, giving the performance for `mismatching` train and test `data`. The phoneme recognition accuracy is shown on second line in Tab. 2. The goal of the re-sampling estimation is to get the phoneme recognition accuracy closer to `matching data` case.

The entropy of the estimator output at given time t is

$$h_t = -\sum_{k=1}^{N} P(q_k|\mathbf{x}_t,\theta) log_2(P(q_k|\mathbf{x}_t,\theta)) \quad (1)$$

where q_k is estimated probability of k^{th} output class of total M classes ($\sum q_k = 1$), $\mathbf{x}_t$ is the input feature vector at time t and θ is set of neural net parameters. A high entropy value means that the outputs are all at some level and the classifier is not able classify input vector. Low entropy value means that the outputs have a peak and rest of the classifier outputs are close to zero – input vector can be clearly classified (although we don't know if correctly).

The average entropy is computed over a *re-sampling estimation* data set. The entropy can be computed on the output of

- band probability estimator. These estimators are not trained very well as they have information from one critical band only. The best trained estimator is for 5^{th} critical band (counting from zero) with cross-validation frame accuracy 38.5%.
- merger probability estimator. This final probability estimator is well trained with cross-validation frame accuracy 68.7%.

The re-sampling coefficient is estimated on 10 sentences from original TIMIT test part. The estimation is done in two phases. In first phase, the average entropy is computed for re-sampling coefficients with large step over whole scale. In the second phase, the re-sampling coefficient steps are smaller with values around the minimum found in first phase. The examples of re-sampling coefficient with corresponding average entropy are shown in Tab. 1.

Finally, the minimum entropy was found together with its re-sampling coefficient. The test data were re-sampled with this coefficient and processed by the system. The recognition accuracy results together with optimal wip are given in Tab. 2. For the `band estimation` experiment, the test data were re-sampled with coefficient obtained from band classifier (1.41) and then processed by already trained system. The data were re-sampled with coefficient 1.56, which was obtained on the merger classifier, for `merger estimation` experiment. The initial experiments results show, that it is possible to estimate re-sampling coefficient which will lead to a good recognition accuracy. These results also suggest to measure the entropy on merger probability estimator.

4 CONCLUSION AND DISCUSSION

Conducted experiments proves TRAP features to be sensitive to different rate of speech in train and test data. The rate was estimated using minimization of entropy at the

re-sampling coefficient	entropy on	
	5^{th} band	merger
1.0	2.92982	1.42335
1.2	2.90826	1.23097
1.4	2.89493	1.10507
1.41	**2.89022**	—
1.43	2.89298	—
1.54	—	1.08185
1.56	—	**1.07700**
1.58	—	1.07924
1.6	2.91637	1.08179
1.8	2.97165	1.09691
2.0	2.99469	1.14223

Table 1: Average entropy as function of re-sampling coefficient

experiment	accuaracy [%]	(wip)
`matching data`	66.4	(8.3)
`mismatching data`	51.3	(-10.0)
`band estimation`	63.2	(14.2)
`merger estimation`	66.1	(9.8)

Table 2: Recognition accuracies [%] with optimal word insertion penalty

output of the classifier - as expected, the estimation after the merger provides significantly better results than estimating the rate at the output of one band-classifier. When the different speech rate is simulated on the TIMIT database, the proposed estimation method works fine and gives us close estimate of re-sampling leading to satisfactory phoneme recognition accuracy.

The future work will aim at testing this estimation on real data, different from the recognizer's training set. We will also work on the estimation of per-speaker rate to improve the recognition accuracy.

REFERENCES

[1] Sharma, S. R.: Multi-stream approach to robust speech recognition, Oregon Graduate Institute of Science and Technology, 1999

[2] Garofolo, J. S., Lanel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L.: DARPA–TIMIT acoustic–phonetic speech corpus, U.S. Department of Commerce, National Institute of Standards and Technology, Computer Systems Laboratory, 1993