Measurement of Complementarity of Recognition Systems

Lukáš Burget

VUT Brno, Faculty of Information Technology
burget@fit.vutbr.cz

Abstract. Combination of different speech recognition systems can be powerful technique to improve recognition performance. The success of these techniques, however, depends on the complementarity of the combined systems. In this paper, a measure of complementarity of different recognition systems is proposed. This measure is based on analysis of similarity of errors made by individual systems. High correlation between the measure and actual performances of combined systems is shown in experiments, which indicates that the measure can be used to select systems suitable for combination. The measure can be computed very efficiently and it can be used even in situations where exhaustive search looking for the set of system optimal for combination would be infeasible.

1 Introduction

In the past, many approaches have been developed to perform speech recognition [1], which differ in feature extraction methods, classification algorithms, methods of model training, and so on. Speech recognition systems based on these different approaches often show important complementarity of their outputs. It has been proved that combination of different systems can be powerful technique to improve recognition performance. The level of success is however limited by the complementarity of systems combined. In this work, we propose a method to measure this complementarity allowing to select such systems whose combination is the most beneficial.

The combination can be performed at different levels. For example, in our experiments, all systems differ only in feature extraction and they could be, therefore, combined directly on feature level, leaving the rest of the system unchanged. In our experiments, however, "hard" outputs of individual recognizers in the form of word (symbol) sequences are combined using technique known as ROVER [2].

Computation of complementarity measure is also based on techniques similar to those used by ROVER. Therefore, ROVER is briefly described in the next section. In section 3, measure of error dependency between **two** recognition systems is developed. In experiments, it is shown that this measure is useful for selection of systems that are good for combination. Measure of complementarity of a **set** of systems is proposed in section 4 and correlation between the measure and actual performances of systems combined using ROVER is shown.

2 ROVER

ROVER (Recognizer Output Voting Error Reduction) [2] is a technique allowing to combine word (symbol) sequences taken as outputs of different recognition systems. Philosophy of this method is illustrated in figure 1. First, alignment of word sequences is performed to find corresponding words over different sequences. In this step, all sequences are merged into one sequence of *correspondence sets*, where each *correspondence set* is a multi-set containing corresponding words one from each individual sequence. In figure 1, *correspondence sets* are represented by columns of words at the output of alignment block. As can be seen in figure 1, there can be no corresponding word from a particular sequence in a *correspondence set*. In such case, *null word* (symbol '-' in figure) is added to the *correspondence set*. In the second step, final symbol sequence is obtained by selecting one word from each *correspondence set* using voting algorithm. In our experiments, simple majority voting is used. Note, that for *correspondence set*, where *null word* is the winning one, no word is output to the final sequence.

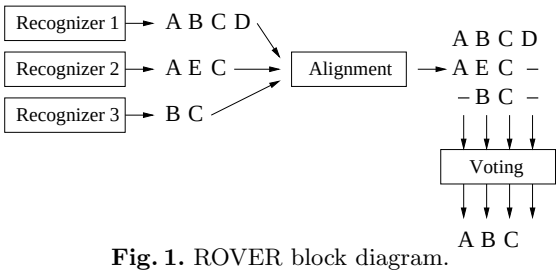

Fig. 1. ROVER block diagram.

The alignment of word sequences is performed similarly to common scoring of speech recognition systems. Dynamic Programing (DP) is used to find such alignment of reference and recognized word sequences that minimizes the cost given by number of insertions, deletions and substitutions. In contrast to alignment used for the scoring, where only two sequences (reference and recognized) are aligned for each utterance, in the case of ROVER, N output sequences corresponding to N combined systems must be aligned. N-dimensional DP would have to be used to obtain optimal alignment, which is very computationally expensive for higher N. An approximation based on iterative application of 2-dimensional DP [2] [3] is used in our experiments.

3 Two system complementarity measure

It was mentioned in section 1 that the improvement of recognition performance given by combination of different systems is limited by the amount of complementarity of systems combined. In our experiments ROVER is used to combine systems at the level of output word sequences. Therefore, we are interested in complementarity encoded in these sequences, which is represented by independency of errors that individual systems make. We will say that two systems

make *dependent error* if both systems make the same error at the same time (e.g. correct word A is recognized by both systems as word B).

3.1 Alignment for identification of error dependency

To identify where two systems make dependent errors, for each utterance from a given set, corresponding output word sequences of both systems are aligned each with other and also with reference word sequence. Alignment of these three sequences is performed in similar manner as ROVER alignment described in section 2. Both output sequences are, however, preferably aligned with the reference sequence. Their alignment with the reference sequence is therefore the same that would be used for scoring. The alignment between output sequences is performed only if more than one alignment with reference sequence having the minimal cost exists. The optimal alignment can be obtained using 3-dimensional DP. However, again, the iterative approximation [3] using 2-dimensional DP is used in our experiments.

3.2 Measurement of two systems error dependency

Once corresponding outputs of two systems are aligned with their references, *dependent errors* can be counted. Let N_{ref} be the total number of words in all reference sequences for the set of utterances used to estimate complementarity measure. Let $N_{dep}(i,j)$ be the total number of *dependent errors* between i^{th} and j^{th} system. We propose the measure called *Dependent Word Error Rate (DWER)* as a measure of error dependency between two systems i and j. The measure is deffined as:

$$DWER(i,j) = N_{dep}(i,j)/N_{ref} \times 100 \quad (1)$$

For a set of systems S and $\forall i,j \in S$, the values of $DWER(i,j)$ form a matrix. We will call this matrix *DWER matrix of set S*. Note, that each value on the matrix diagonal $DWER(i,i)$, which is the ordinary WER for the system i, is the highest value in the corresponding row and column.

3.3 Experimental setup

Speech data from TI Connected Digits database [4] were used for both training and testing of all recognition systems. Limited number of clean speech utterances were selected for training (616 utterances from 4 male and 4 female speakers). Four types of noise (subway, car, exhibition, babble) from AURORA2 TI Digits database [5] were artificially added to speech data at SNR level 20dB and 10dB. The same 616 utterances were used to create data for all noisy conditions. Together $616 \times (1 + 4 \times 2) = 5544$ utterances were used for training. Test data were prepared in a similar manner. Here, 912 utterances from 12 male and 12 female speakers were used. Together $912 \times (1 + 4 \times 2) = 8208$ utterances were used for testing. Nine recognition systems were trained, each using different feature extraction method. The following feature extraction method were used: **BSL** - 15

Table 1. Word Error Rates of individual recognizers.

System	POW	DA4	30B	ENG	BLS	15B	LPCC	DA1	NOE	ROVER 9
WER [%]	2.90	2.91	2.99	3.00	3.04	3.14	3.36	3.51	3.58	2.59

Mel Frequency Cepstral Coefficients (MFCC) [6] augmented with their first and second order derivatives (delta and double-delta), filter bank applied on magnitude spectrum, 23 bands in Mel filter bank, 25 ms window length, 10 ms frame rate, 5 frames delta and delta-delta window, frame energy is represented by C0 coefficient. **LPCC** - 15 LPCC augmented with their derivatives (LPC order 15, other parameters similar to BSL features). The name BSL stays for "baseline", since all seven remaining feature extraction methods are only modifications of BSL methods and always only one of their parameters is changed. In the following list, only the changed parameter of BSL features is described: **DA1** - delta and delta-delta window is 3 frames; **DA4** - delta and delta-delta window is 9 frames; **B15** - 15 bands are used in filter bank; **B30** - 30 bands are used in filter bank; **POW** - filter bank is applied on power spectrum; **NOE** - only coefficients C1 to C14 are used; **ENG** - frame energy replaces C0 coefficient.

Except the feature extraction part, all recognition systems are the same. Continuous HMMs are used with output probability density function modeled by Gaussian mixture (3 components). Whole word models with left-to-right topology (16 states for digits, 3 states for silence) are used. Table 1 shows WER of all individual recognition systems. In the last column, there is WERs for ROVER combination of all nine systems. The performance of ROVER is significantly better than performance of any individual system.

3.4 Analysis of DWER matrix

In our experiments, we will be interested in the correlation between the actual recognition performance of combined system and proposed complementarity measure. Therefore the test data are also used to derive *DWER matrix*.

For our set of nine systems, the estimate of *DWER matrix* defined by equation 1 is shown in table 2. In the table, it can be directly observed, that values in the row and column corresponding to system DA4 are considerably smaller than other values. These lower values of DWER indicate high complementarity of DA4 system with all other systems. More over, among the systems in our set, DA4 system has second lowest WER. Therefore, it will be the hot candidate for combining. Second system that seems to be quite complementary to other systems is LPCC.

Complementarity of both systems DA4 and LPCC is probably even more visible on figure 2, which is graphical representation of *DWER matrix*. Bright rows and and columns corresponding to DA4 and LPCC systems represent low DWER values. In opposite, we can see darker block representing DWERs between systems POW, 30B, ENG and BSL, indicating higher error dependency, which is (as we expect) caused by their lower complementarity.

Table 2. DWER matrix for set of nine systems.

System	POW	DA4	30B	ENG	BSL	15B	LPCC	DA1	NOE
POW	2.9	1.5	2.0	2.0	2.2	1.9	1.6	1.8	1.9
DA4	1.5	2.9	1.4	1.5	1.5	1.4	1.3	1.4	1.6
30B	2.0	1.4	3.0	1.9	2.3	1.7	1.5	1.9	1.7
ENG	2.0	1.5	1.9	3.0	2.0	1.8	1.6	1.8	1.8
BSL	2.2	1.5	2.3	2.0	3.0	1.8	1.5	1.9	1.8
15B	1.9	1.4	1.7	1.8	1.8	3.1	1.6	1.6	1.8
LPCC	1.6	1.3	1.5	1.6	1.5	1.5	3.4	1.5	1.5
DA1	1.8	1.4	1.8	1.8	1.9	1.6	1.5	3.5	1.8
NOE	1.9	1.6	1.7	1.8	1.8	1.8	1.5	1.8	3.6
Avg.	1.66	1.29	1.59	1.61	1.67	1.50	1.35	1.52	1.53

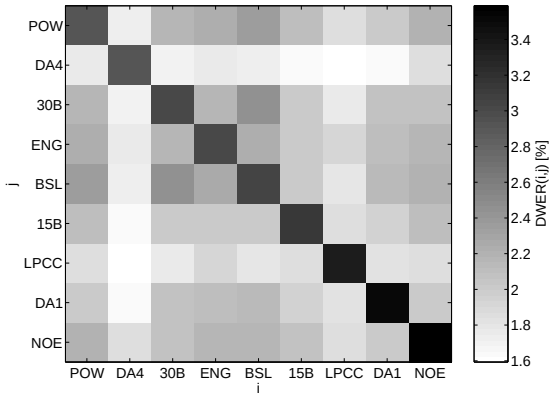

Fig. 2. DWER matrix for set of nine systems.

3.5 Redundancy of a system in the system set

As an objective measure of one system complementarity with all other systems in the set, we propose to simply average values in *DWER matrix* column (or row) corresponding to the system. Ordinary WERs of the systems (diagonal values) are excluded from averaging. These column averages can be seen in the last rows of table 2. In the table, we observe that the lowest values indicating high complementarity with other systems correspond to systems DA4 and LPCC, which is in agreement with our previous findings. In opposite, the highest value indicating low complementarity corresponds to BSL system. This is a natural finding, because all other systems (except of LPCC) use features which are derived from BSL features by modifying only one of its parameters.

The proposed measurement of one system complementarity with all other systems is verified in the experiment, where only eight of nine systems are combined using ROVER. Here, we are interested in performance degradation when excluding one particular system from combination. In the table 1, we saw that WER of ROVER combination of all nine systems is 2.59%. Table 3 shows com-

bined system WERs depending on which system is excluded from combination. The highest degradation is caused by omitting system DA4, followed by systems LPCC, which were "marked" as two systems most complementary to other systems according to proposed complementarity measure. In opposite, three least complementary systems according to the measure are BSL, POW and ENG. As can be seen in the table 3, performance of ROVER even *improves* when excluding one of these three systems from combination.

Table 3. ROVERing 8 of 9 systems. Some combinations of eight systems perform even better than the combination of all nine systems with WER of 2.59%. WERs of such combined systems are indicated by bold values.

Excluded system	POW	DA4	30B	ENG	BSL	15B	LPCC	DA1	NOE
ROVER WER	**2.49**	2.66	2.61	**2.58**	**2.54**	2.61	2.63	2.62	2.61

4 Complementarity measure for set of systems

In the previous section, we have shown some connection between complementarity of recognition systems, their suitability for system combination and DWER measure corresponding to these systems. Values from *DWER matrix* were used to make a decision which systems from a given set are complementary to others and which are redundant for system combination. However, it would be practical to have a measure assigning a single value to a system **set**, that would say how the systems from the set are good for combination. In the ideal case, this measure would allow to select the subset of a large set of systems whose combination would lead to lowest WER. Complementarity measure for a set of systems is proposed in this section and the correlation between proposed measure and actual WER of combined system is shown in experiments.

4.1 Average Dependent Word Error Rate (ADWER)

In the previous section, average of DWER matrix column was used as a measure of one system complementarity with all other systems in the given set. As a natural extension, we propose to simply average all values from *DWER matrix* to obtain measure of overall complementarity among systems in a set S ($|S|$ denotes number of systems in this set):

$$ADWER(S) = \sum_{i \in S} \sum_{j \in S} DWER(i,j)/|S|^2 \tag{2}$$

4.2 Experimental setup

In the experiments with system set complementarity measure, the same training and testing data described in section 3.3 are used. Test data are again used for estimation of DWER matrix. All individual systems are identical to those described in section 3.3, in addition two systems using features, which are again derived from BSL features, were included to the system set, **W15** with frame window length 15ms and **W35** with the frame window length 35ms.

4.3 Correlation between combined system WER and system set complementarity measure

From the set of eleven systems, all subsets consisting of three and eight systems were combined using ROVER and corresponding WERs were evaluated. Combination of three and eight systems was chosen to show how complementarity measure is correlated with combined system WER for combinations of only few (three) and larger number (eight) of systems.

Figure 3 shows correlation between WER of combined system (X axis) and average of WERs of corresponding individual systems (axis Y). Each dot, cross and big cross in the figure corresponds to one combination of three, eight and eleven systems respectively. Figure 3 shows that no significant correlation can be observed and, therefore, we can conclude that WERs of individual systems are not important for selection of systems suitable for combination.

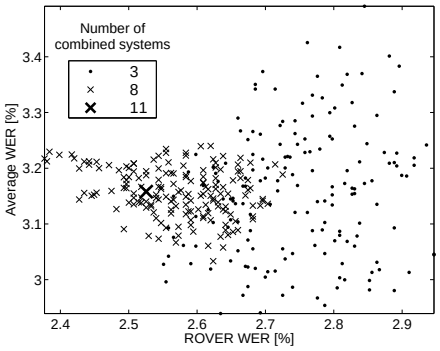

Fig. 3. Correlation between WER average and ROVER WER.

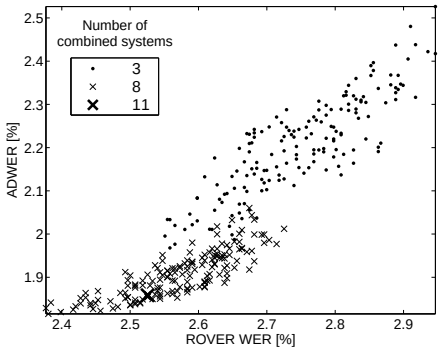

Fig. 4. Correlation between ADWER and ROVER WER.

In the following experiments, we will see a correlation between proposed system set complementarity measure and corresponding combined system WER.

Figure 4 shows the correlation between combined system WER and corresponding *Average Dependent Word Error Rate (ADWER)* measure computed according to equation 2. For both combinations of three and eight systems, visible correlation is observed between ADWER measure and combined system WER.

It can be seen in figure 4 that dots representing combinations of three systems and crosses representing combinations of eight systems are concentrated around two separate lines. Therefore, values of ADWER measure can not be compared for two sets with different number of systems. In other words, first, we must know how many systems we want to combine and then we can use ADWER measure to choose which systems will be good for combination.

5 Discussion and conclusions

Combination of different systems can be a powerful technique to improve recognition performance. The success of these techniques is, however, contingent on complementarity of combined systems. Given a set of N systems, one way to determine the subset of systems most suitable for combination is to exhaustively evaluate recognition performance for all possible system combinations. In the case of ROVER-like combination of system output sequences, training and recognition must be performed only once for each of N systems. Then, however, ROVER-like technique must be applied for each combination of N systems, which may be not feasible for large values of N. From this point of view, combination on the feature level is even worse case. Here, also training and recognition must be performed for each combination of N systems, which increases the whole evaluation time in order of magnitudes. For this reason, we have proposed measure of recognition systems complementarity, which are based on measurement of error dependency of individual system outputs. First, method for measuring complementarity of two systems was proposed. This measure can be computed very efficiently even for large set of systems. Training and recognition must be performed only once for each of N systems, then technique similar to ROVER is used to measure complementarity only for each pair of systems. Simple averaging of the measure is used as an extension allowing to measure the complementarity of a system subset. Correlation between the measure and actual performance of combined systems was shown in experiments, which indicates that this measure can be advantageously used to select systems suitable for combination.

6 Acknowledgments

This research has been partially supported by Grant Agency of Czech Republic under project No. 102/02/0124 and by EC project Multi-modal meeting manager (M4), No. IST-2001-34485.

References

1. B. Gold and N. Morgan, *Speech and audio signal processing*, John Wiley & Sons, 2000.
2. J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. IEEE Workshop on automatic speech recognition and understanding*, 1997.
3. L. Burget, "Measurement of complementarity of recognition systems," Tech. Rep., Brno University of Technology, Faculty of Information Technology, 2003.
4. R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP'84*, 1984.
5. D. Pearce H. G. Hirsch, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Automatic Speech Recognition: Challenges for the Next Millennium*. ISCA ITRW ASR2000, Paris, France.
6. S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech & Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.