Measurement of complementarity of recognition systems

Lukáš Burget

Brno University of Technology, Faculty of Information Technology
Brno, Božetěchova 2, 612 66, Czech republic
burget@urel.fee.vutbr.cz

July 2, 2004

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, measures of complementarity of different recognition systems are proposed. These measures are based on analysis of similarity of errors made by individual systems. High correlation between these measures and actual performances of combined systems is shown in experiments, which indicates that these measures can be used to select systems suitable for combination. The measures can be computed very efficiently and they can be used even in situations where exhaustive search looking for the set of systems optimal for combination would be infeasible.

1 Introduction

In the past, many approaches have been developed to perform speech recognition, which differ in feature extraction method (MFCC [3], PLP [7], TRAPS [8] [9]), classification algorithm (HMM [1] [2], Hybrid ANN-HMM [12]), used model (different HMM types and topologies), method of model training (Maximum Likelihood (ML) [1], discriminative training (MMI [15], MCE [16])), and so on. It is not possible to say, which approach is the right one. For example, in the case of feature extraction, it is not exactly known which information should be extracted from speech. Moreover, attempt to preserve one information often leads to loss of another (e.g. resolution in the time vs. resolution in the frequency). 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 method and they could be,

therefore, combined directly on feature level, leaving the rest of the system unchanged. In this case, individual feature streams could be combined into one stream using some technique (such as PCA, LDA [6] [11] [10], Tandem [13] [14]) preserving the important information encoded in the original streams. In our experiments, however, "hard" outputs of individual recognizers in the form word (symbol) sequences are combined using technique known as ROVER (Recognizer Output Voting Error Reduction) [17]. Measures of complementarity are also based on comparing output word sequences, which however does not mean, that this measures are not meaningful for other methods of system combination.

Computation of complementarity measures is also based on techniques similar to those used by ROVER. Therefore, ROVER is briefly described in the next section. In section 4, measures of error dependency between **two** recognition systems are developed. In experiments, it is shown that these measures are useful for selection of systems good for combination. Measures of complementarity of **set** of systems are proposed in section 5 and correlation between these measures and actual performances of systems combined using ROVER is shown.

2 Terminology

In following text, term *system* will be used to denote individual speech recognition system. Terms *system output* or *output sequence* will denote sequence of words recognized by the system. Term *combined system* will be used for combination of individual systems. ROVER is used in our experiments for system combination, therefore, *combined system output* is obtained as ROVER combination of output word sequences of individual recognition systems. Term *system set* will be used to denote set of individual systems available for combination. Where only several systems from currently used system set are combined, term *system subset* will be used.

3 ROVER

ROVER (Recognizer Output Voting Error Reduction) [17] 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 is performed to find corresponding words over different output sequences. In this step, outputs of all recognizers are merged into one sequence of *correspondence set*, where each *correspondence sets* is a multi-set [1] containing corresponding words one from each recognizer output. In figure 1, *correspondence sets* are represented by columns of words on the output of alignment block. As can be seen in figure 1, there can be no word in a particular sequence corresponding to 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.

[1] Set where multiple occurrences of the same element are allowed

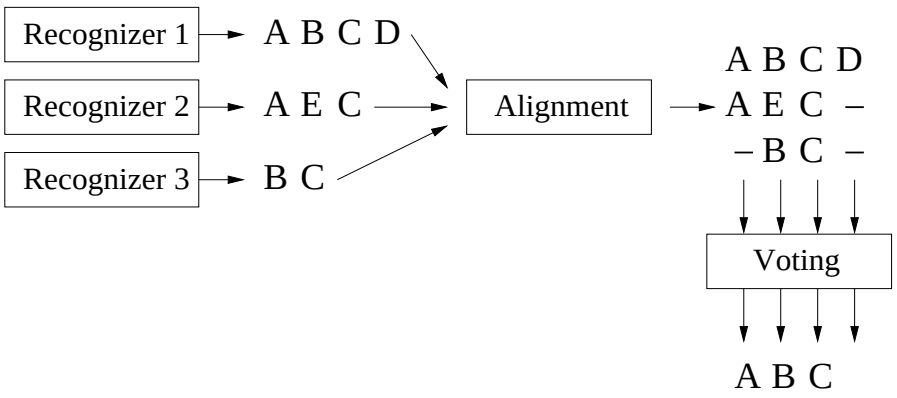


Figure 1: ROVER method block diagram.

3.1 ROVER Alignment

Merging of individual word sequences into one sequence of *correspondence sets* is performed iteratively. Initially, first two sequences are aligned to produce *correspondence sets* each having only two elements. The alignment is performed the same way that is commonly used for scoring performances of speech recognition systems. Reference word sequence is aligned with the recognized one, to allow for counting of insertions, deletions and substitutions. Such alignment, that minimize the total coast is found using Dynamic Programing. In our implementation, the cost of each deletion and insertion is 3, the cost of a substitution is 4 and cost of a correct word is 0. Of course, in ROVER alignment we do not have any reference sequence, however, assuming the first sequence being the reference will allow us to use terms: insertion and deletion in the following examples.

In the example from figure 1, *correspondence set* sequence created in the first iteration is:

$$\begin{array}{r} 1st\ sequence \\ 2nd\ sequence \end{array} \begin{pmatrix} A \\ A \end{pmatrix}, \begin{pmatrix} B \\ E \end{pmatrix}, \begin{pmatrix} C \\ C \end{pmatrix}, \begin{pmatrix} D \\ - \end{pmatrix}$$

and cost for this alignment is: $4 + 3 = 7$ corresponding to one substitution of the word E for B and one deletion of the word D.

In each next iteration, next word sequence is aligned with the *correspondence sets* obtained in previous iteration. The alignment is performed the same way as the alignment of first two sequences, however, sequence of *correspondence sets* now serve as the reference and cost must be always computed with respect to all words in each *correspondence set*. In the example from figure 1, *correspondence set* sequence created in the second (last) iteration is:

$$\begin{array}{r} 1st\ sequence \\ 2nd\ sequence \\ 3rd\ sequence \end{array} \begin{pmatrix} A \\ A \\ - \end{pmatrix}, \begin{pmatrix} B \\ E \\ B \end{pmatrix}, \begin{pmatrix} C \\ C \\ C \end{pmatrix}, \begin{pmatrix} D \\ - \\ - \end{pmatrix}$$

and cost for this alignment is: $2 \times 3 + 4 + 3 = 13$ corresponding to two deletions of the word A, one substitution of B for E and one deletion of D.

Note that this iterative method of alignment of multiple sequences is not the optimal one and the order in which individual sequences are aligned is important.

In our experiments, output of systems that performs the best [2] are aligned first, outputs of systems with poorer performance are added later. This has been experimentally shown to be a reasonable suboptimal solution. N-dimensional Dynamic Programing would have to be used to obtain optimal alignment, where N is the number of sequences aligned. However, such alignment would be very computationally expensive for higher N.

4 Measures of complementarity of two recognition system outputs

It was mentioned in section 1 that the improvement of recognition performance given by combination of different systems is limited by 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 distinguish two types of error dependency. We will say that two systems make *simultaneous error* if both systems make error at the same time. Both systems can, however, make different errors (e.g. correct word A is recognized by first system as B and by second system as C). We will say that two systems make *dependent error* in the special case where both systems make the same error.

In this section, measures of complementarity of two systems based on counting *simultaneous and dependent errors* are proposed. Extension for measuring complementarity of a whole set of systems will be proposed in section 5. Measures of complementarity of two systems are estimated on a selected set of utterances in the following steps:

- For each utterance, output sequences of both systems are obtained.
- Each pair of sequences is aligned with corresponding reference sequence according to algorithm described in section 4.1
- For each pair of sequences, *simultaneous and dependent errors* are counted (see section 4.2).
- Counts of *simultaneous and dependent errors* are used to compute complementarity measures proposed in section 4.3

4.1 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 with reference word sequence. Alignment is performed in similar manner as ROVER alignment described in section 3.1. Output sequence of one system is aligned with reference sequence first. However, when the second output sequence is added the alignment is performed with respect to words only from reference sequence, and output of first system is taken into account, only if more than one alignment with reference sequence having the minimal cost is available. This is best illustrated on following the example:

[2] In terms of Word Error Rate of individual systems.

Let the reference sequence be only one word C. Both systems tend to insert words so that output sequences of the first and the second system are: C, X, X, X, Z and and X, X, X, C, Z respectively. These two sequences will be referred as *sequence 1* and *sequence 2*. The alignment with minimal cost that would be used by ROVER for all three sequences is:

$$\begin{array}{l} reference \\ sequence\ 1 \\ sequence\ 2 \end{array} \left(\begin{array}{c} C \\ C \\ - \end{array}\right), \left(\begin{array}{c} - \\ X \\ X \end{array}\right), \left(\begin{array}{c} - \\ X \\ X \end{array}\right), \left(\begin{array}{c} - \\ X \\ C \end{array}\right), \left(\begin{array}{c} - \\ Z \\ Z \end{array}\right)$$

In this case, alignment of words from *sequence 1* and words from reference is the same that would be used for scoring system performance (see section 3.1). For identification of error dependency, we would like to have alignment of *sequence 2* and reference sequence also the same as the one used for scoring. However, *sequence 2* is not aligned with reference in such manner in ROVER alignment (word C from *sequence 2* is not aligned with word C from reference). For this reason, in alignment used for identification of error dependency, *sequence 2* is preferably aligned to the reference sequence. The following alignment is obtained for sequences from our example:

$$\begin{array}{l} reference \\ sequence\ 1 \\ sequence\ 2 \end{array} \left(\begin{array}{c} - \\ - \\ X \end{array}\right), \left(\begin{array}{c} - \\ - \\ X \end{array}\right), \left(\begin{array}{c} C \\ C \\ C \end{array}\right), \left(\begin{array}{c} - \\ X \\ - \end{array}\right), \left(\begin{array}{c} - \\ X \\ - \end{array}\right), \left(\begin{array}{c} - \\ X \\ - \end{array}\right), \left(\begin{array}{c} - \\ Z \\ Z \end{array}\right)$$

When aligning *sequence 2*, words form *sequence 1* are initially ignored and the word C is therefore aligned with the word C from the reference sequence. The word Z can be, however, aligned with the same cost with any *null word* following C in the reference sequence. Here, the words from *sequence 1* are also taken into account and the secondary cost minimization leads to the alignment of the words Z from *sequence 1* and *sequence 2*.

Note that the order in which two output sequences are aligned is again important. Following example demonstrate the case where different order results in different alignments. If sequence X is aligned with reference sequence first and then sequence Y is added, following optimal alignment is obtained:

$$\begin{array}{l} reference \\ sequence\ X \\ sequence\ Y \end{array} \left(\begin{array}{c} B \\ B \\ B \end{array}\right), \left(\begin{array}{c} C \\ C \\ - \end{array}\right), \left(\begin{array}{c} A \\ B \\ B \end{array}\right)$$

Opposite order of processing sequences Y and X can lead to the following wrong alignment:

$$\begin{array}{l} reference \\ sequence\ Y \\ sequence\ X \end{array} \left(\begin{array}{c} B \\ B \\ B \end{array}\right), \left(\begin{array}{c} C \\ B \\ C \end{array}\right), \left(\begin{array}{c} A \\ - \\ B \end{array}\right)$$

Until we see sequence X we do not know that it is better to align second word B from sequence Y with word A from reference and alignment with word C that has the same cost can be chosen. Adding sequence X already does not affect this wrong decision.

$$\begin{array}{l} reference \\ sequence\ Y \\ sequence\ X \end{array} \left(\begin{array}{c} B \\ B \\ B \end{array}\right), \left(\begin{array}{c} C \\ B \\ C \end{array}\right), \left(\begin{array}{c} A \\ - \\ B \end{array}\right)$$

Note that the optimal alignment can be obtained using 3-dimensional Dynamic Programing.

4.2 Counting simultaneous and dependent errors

Once corresponding outputs of two systems are aligned with their references, *simultaneous and dependent errors* can be counted. The following example demonstrates alignment of sequences with two *simultaneous errors* where words A and D are incorrectly recognized by both systems. Moreover, in the case of word D, both systems make the same error (words deleted) and therefore this error is also *dependent error*.

$$\begin{array}{c} reference \\ output\ 1 \\ output\ 2 \end{array} \begin{pmatrix} A \\ E \\ F \end{pmatrix}, \begin{pmatrix} B \\ B \\ B \end{pmatrix}, \begin{pmatrix} C \\ G \\ C \end{pmatrix}, \begin{pmatrix} D \\ - \\ - \end{pmatrix}$$

For measuring error dependency, we are not interested in the cases where only one system makes error (word C is incorrectly recognized only by first system).

4.3 Measurement of error dependency between two systems

Let N_{ref} be the total number of words in all reference sequences for the set of utterances used to estimate complementarity measures. Let $N_{sim}(i,j)$ and $N_{dep}(i,j)$ be the total number of *simultaneous errors* and *dependent errors* between i^{th} and j^{th} system respectively. We propose the following measures of error dependency between two systems:

4.3.1 Lower Bound Word Error Rate (LBWER)

for two systems i and j is defined as ratio between the number of *simultaneous errors* and the overall number of words in the set of utterances:

$$LBWER(i,j) = \frac{N_{sim}(i,j)}{N_{ref}} \times 100 \tag{1}$$

We can also regard this measure as error rate of such system combining outputs of two recognizer that always select (using an ideal confidence measure) the correct word for all the cases where only one recognizer makes error (therefore the name Lower Bound WER).

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

4.3.2 Dependent Word Error Rate (DWER)

for two systems i and j is defined as ratio between the overall number of *dependent errors* and the number of words in our set of utterances:

$$DWER(i,j) = \frac{N_{dep}(i,j)}{N_{ref}} \times 100 \tag{2}$$

DWER matrix of a system set is defined in the same manner as *LBWER matrix*. Note, that values on the *DWER matrix* diagonal are again ordinary WERs of individual systems.

4.4 Properties of error dependency measures

If a set of at least three systems has diagonal *LBWER matrix*, ROVER combination of these systems based on majority voting must result in zero WER. Systems make no *simultaneous errors* in such case, and therefore a single system making an error is always outvoted by all others. Note, that this does not have be true for a set of systems with diagonal *DWER matrix*. On the other hand, *dependent errors* measured by DWER can be seen as the worse variants of *simultaneous errors*, since any time systems make *dependent error*, we need even more correct answers to outvote the error. We can, therefore, intuitively expect, that system set, in order to be good for combination, must generally have small values out of *LBWER matrix* and *DWER matrix* diagonals (but perhaps also values on diagonals representing ordinary WERs should be small, since performance of individual system is also important).

Both *LBWER* and *DWER matrices* are, however, not directly related to performance of combined system. It can be proved on the following example showing two sets of systems with identical *LBWER* and *DWER matrix* where combination of systems from each set leads to different results.

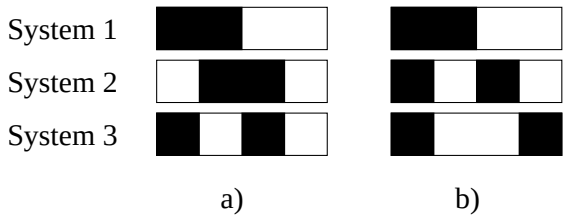


Figure 2: Different systems with the same DWER matrix.

Table 1: DWER matrix for both systems from figure 2.

	Sys. 1	Sys. 2	Sys. 3
System 1	50	25	25
System 2	25	50	25
System 3	25	25	50

Figure 2a represents a set of three systems, where each row bar corresponding to one system shows portion of correctly recognized words (white area) and incorrectly recognized words (black area). Overlapping parts of black areas of two systems correspond to *simultaneous error* between the systems. Let us assume that all *simultaneous errors* are also *dependent errors* in this example. Therefore *LBWER matrix* for this system set, which is shown in table 1, will be identical to *DWER matrix*. WERs of all individual systems are 50% (values in matrix diagonal), *dependent errors* for all pairs of systems are 25% (values out of diagonal). Majority voting based combination of systems from figure 2a would result in WER of 75%, since all systems vote for correct word only for the portion corresponding to last quarter of row bars. In all other cases, there are always two systems making error outvoting the correct one. The combined systems therefore perform even worse than the individual system. On the other hand, system combination not based only on

majority voting can, for example, choose the output word using a word confidence measure that in ideal case always prefers the correct word. Such combination can still result in WER 0%, since, there is always at least one system producing the correct word in figure 2a.

Figure 2b represents a set of systems that have *LBWER and DWER matrices* identical to those obtained for system set from figure 2a. In this case however, majority voting based combination would result in better WER of 25% and in opposite, the mentioned word confidence measure based system combination cannot have WER better then 25% since all systems make error for the portion corresponding to first quarter of row bars. The difference is caused by the triple error dependency (three systems make dependent error) that is already not captured in *LBWER and DWER matrices* representing only dependences between pairs of systems.

Although it was shown that there is no direct relationship between performance of combined systems [3] and values from LBWER and DWER matrices, we can still expect that there is certain correlation between them. In the next section, experimental setup will be described, where recognition systems using different feature sets are combined by ROVER. LBWER and DWER matrices of these systems will be analyzed and we will observe that LBWER and DWER measures can be useful for selection of systems good for combination. In section 5, complementarity measures for a whole set of systems are proposed, which are based on the values from *LBWER and DWER matrices*. Correlation between these measures and performance of ROVER combining the system set is shown.

4.5 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, 4 noises used for training and four unseen noises (train station, airport, restaurant, street) were added to test data. Additionally, SNR 0dB condition was generated for both seen and unseen noises. Together $912 \times (1 + 8 \times 3) = 22800$ 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 Mel Frequency Cepstral Coefficients [3] 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)

[3] At least for the combination techniques mentioned in the example.

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 instead of 5 frames
- **DA4** - delta and delta-delta window is 9 frames instead of 5 frames
- **B15** - 15 bands are used in filter bank instead of 23 bands
- **B30** - 30 bands are used in filter bank instead of 23 bands
- **ENG** - frame energy is computed as replacement for C0 coefficient
- **POW** - filter bank applied on power spectrum instead of magnitude spectrum
- **NOE** - only coefficients C1 to C14 are used (no C0 or frame energy)

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 mixture components). Whole word models with left-to-right topology (16 states for digits, 3 states for silence) are used.

Names of feature extraction methods will be used also to distinguish individual systems in following text. For example, system using BSL features will be referred as BSL system.

Table 2: Word Error Rates of individual recognizers.

Condition	**Clean**	**Seen noises**			**Unseen noises**			**Seen**
SNR level	-	**20dB**	**10dB**	**0dB**	**20dB**	**10dB**	**0dB**	**cond.**
System:								
POW	1.11	1.70	4.55	48.50	1.50	3.70	37.03	2.90
DA4	1.34	1.58	4.65	48.67	1.45	3.63	36.34	2.91
30B	1.76	1.62	4.68	52.55	1.57	3.77	40.38	2.99
ENG	1.37	1.69	4.72	44.11	1.63	4.12	35.97	3.00
BLS	1.37	1.75	4.74	51.18	1.58	3.77	38.51	3.04
15B	1.63	1.63	5.03	51.66	1.54	4.20	40.94	3.14
LPCC	1.44	1.62	5.59	44.50	1.64	4.41	29.97	3.36
DA1	1.89	2.06	5.39	54.87	1.80	4.30	44.73	3.51
NOE	3.59	1.71	5.47	58.52	1.97	4.80	48.66	3.58
ROVER 9	1.14	1.41	4.14	49.55	1.35	3.38	37.38	2.59

Table 2 shows WER of all individual recognition systems for different levels of SNR for both seen and unseen conditions. All values in the table for seen and unseen noises are averaged accuracies for four seen or four unseen types of noise. In the experiments, we will need single value representing the system performance, with respect to which we can look for the optimal system combination. For this purpose, we will use WER evaluated on subset of test data containing: clean data and data corrupted by seen noises with SNR 20dB and 10dB. This data subset will be referred as *seen conditions test data*. WERs for individual recognition systems evaluated on

this subset can be seen in the last column of table 2. In the last row, there are WERs for ROVER combination of all nine systems. The overall performance of ROVER is generally better than performance of any individual system, however, for certain conditions (clean speech and SNR 0dB) some system are able to even outperform ROVER.

4.6 Analysis of LBWER and DWER matrices

Seen conditions test data are also used to derive *LBWER* and *DWER matrices*. Here, one could object that test data should not be used for estimation of complementarity measures based on *LBWER and DWER matrices*. In our experiments, we will see the correlation between proposed complementarity measures and the actual recognition performance of combined system. We will be, however, interested in the true correlation observed for the ideal case, where measures are estimated and system is evaluated on the same data. We can also consider *seen conditions test data* to be an evaluation set that is only intended to find the best combined system. Another question is, how error dependency statistics estimated on this data set will generalize for other test data. For this purpose, we can look at results obtained for unseen noises, which are not used for estimation of *LBWER and DWER matrices*

For our set of nine systems, the estimate of *LBWER matrix* defined by equation 1 is shown in table 3. Values in the matrix diagonal are ordinary WERs of individual systems from the last column of table 2. Although *LBWER matrix* should be symmetric, we can see that corresponding values slightly differ in table 3. In the section 4.1, suboptimal alignment used in our experiments for estimation of LBWER and DWER was described and the importance of the order in which output sequences of two systems are aligned with reference sequence was noticed. Each two corresponding values in table 3 correspond to these two different alignment orders. The differences between the corresponding values are, however, very small, which proves the proper functionality of the suboptimal alignment method used. *DWER matrix* with similar properties defined by equation 2 can be found in table 4.

In both tables 3 and 4, 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 LBWER and 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 3, which is graphical representation of *LBWER matrix*. Bright rows and and columns corresponding to DA4 and LPCC systems represent low LBWER values. In opposite, we can see darker block representing LBWERs between systems POW, 30B, ENG and BSL, indicating higher error dependency between these systems, which is (as we expect) caused by their lower complementarity. Figure 4 showing similar graphical representation of *DWER matrix*, is visually almost identical with figure 3.

Dependent errors were defined as a special case of *simultaneous errors*. Table 5 shows how many percent of *simultaneous errors* are also *dependent errors* for each pair of systems. All values in the diagonal are 100%, which corresponds to the fact that if two same systems are compared, all their errors are *simultaneous errors* and at the same time also *dependent errors*. For any pair of systems, most of *simultaneous*

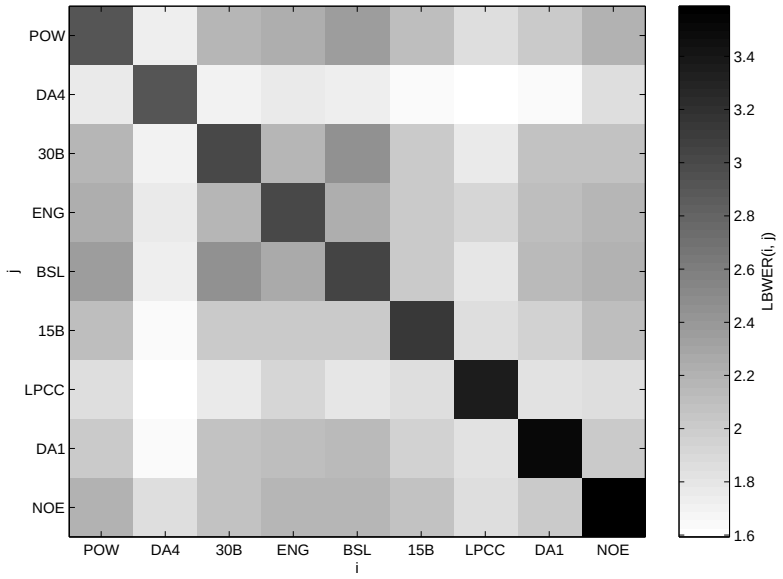


Figure 3: LBWER matrix for set of nine systems.

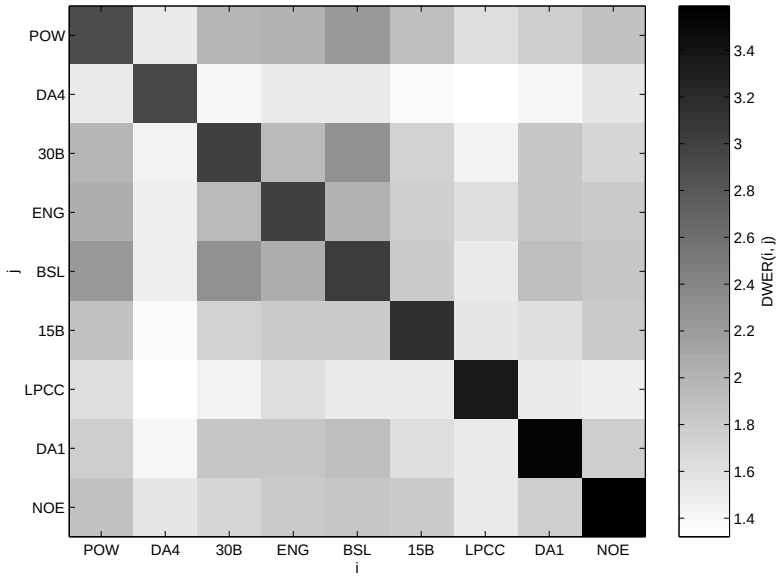


Figure 4: DWER matrix for set of nine systems.

Table 3: LBWER matrix for set of nine systems.

System	POW	DA4	30B	ENG	BSL	15B	LPCC	DA1	NOE
POW	2.90	1.75	2.17	2.23	2.36	2.11	1.85	2.01	2.20
DA4	1.75	2.92	1.72	1.76	1.74	1.65	1.60	1.65	1.85
30B	2.18	1.71	3.00	2.17	2.46	2.01	1.77	2.09	2.09
ENG	2.22	1.75	2.16	3.00	2.25	2.03	1.92	2.10	2.17
BSL	2.36	1.74	2.46	2.26	3.04	2.03	1.80	2.14	2.19
15B	2.11	1.64	2.01	2.02	2.02	3.14	1.86	1.94	2.10
LPCC	1.85	1.59	1.77	1.91	1.80	1.86	3.36	1.81	1.86
DA1	2.01	1.65	2.09	2.10	2.14	1.94	1.82	3.52	2.03
NOE	2.19	1.85	2.09	2.17	2.18	2.09	1.87	2.02	3.59
Avg.	1.85	1.52	1.83	1.85	1.88	1.75	1.61	1.75	1.83

Table 4: DWER matrix for set of nine systems.

System	POW	DA4	30B	ENG	BSL	15B	LPCC	DA1	NOE
POW	2.90	1.51	1.97	2.02	2.22	1.89	1.62	1.76	1.88
DA4	1.52	2.92	1.43	1.50	1.50	1.36	1.33	1.43	1.56
30B	1.98	1.43	3.00	1.93	2.30	1.73	1.46	1.85	1.71
ENG	2.03	1.50	1.93	3.00	2.03	1.77	1.62	1.83	1.79
BSL	2.22	1.50	2.30	2.04	3.04	1.78	1.51	1.91	1.84
15B	1.89	1.36	1.73	1.78	1.79	3.14	1.56	1.64	1.79
LPCC	1.61	1.32	1.46	1.61	1.51	1.53	3.36	1.51	1.48
DA1	1.77	1.42	1.84	1.84	1.89	1.63	1.51	3.52	1.75
NOE	1.88	1.55	1.70	1.79	1.84	1.79	1.51	1.75	3.59
Avg.	1.66	1.29	1.59	1.61	1.67	1.50	1.35	1.52	1.53

errors are also *dependent errors* (between 79.7% and 94.2%). We can, therefore, expect that measurement of complementarity based on LBWER will not be too different from that based on DWER. Still there is visible difference in the percentage of *dependent errors* for different pairs of systems, which justifies investigating both kinds of complementarity measurements.

4.7 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 *LBWER or DWER matrix* column (or row) corresponding to the system. Ordinary WERs of the systems (values on the diagonal) are excluded from averaging [4]. These column averages can be seen in

[4]Including system's own WERs in averaging can be seen as additional "bonus" for systems with low WER. That may be also important while selecting systems for their combination. The effect of including or excluding WERs from complementarity measure computation will be demonstrated in experiments described in section 5.

Table 5: Percentage of *dependent errors* in *simultaneous errors*.

System	POW	DA4	30B	ENG	BSL	15B	LPCC	DA1	NOE
POW	100.0	86.7	90.6	90.6	94.2	89.7	87.5	87.6	85.5
DA4	86.7	100.0	83.1	85.3	86.2	82.0	83.0	86.6	84.1
30B	91.0	83.7	100.0	89.3	93.5	85.9	82.6	88.4	81.9
ENG	91.4	85.5	89.3	100.0	90.3	87.5	84.7	87.2	82.4
BSL	94.2	86.0	93.2	90.2	100.0	87.8	83.7	89.0	83.9
15B	89.5	83.0	86.1	88.0	88.2	100.0	84.0	84.5	85.5
LPCC	86.9	82.9	82.5	84.4	83.8	82.2	100.0	83.0	79.7
DA1	87.9	86.3	88.0	87.9	88.6	84.1	83.1	100.0	86.4
NOE	85.8	83.9	81.6	82.7	84.2	85.6	80.7	86.5	100.0

Table 6: 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 in the table.

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

the last rows of tables 3 and 4. In both tables, we observe that the lowest values indicating high complementarity with other systems corresponds 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 an interesting and 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 2, we saw that WER of ROVER combination of all nine systems is 2.59%. Table 6 shows combined 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 distinguished as two systems most complementary to other systems according to proposed complementarity measures. In opposite, three least complementary systems according to the measures are BSL, POW and ENG. As can be seen in the table 6, performance of ROVER even improves when excluding one of these three systems from combination.

In the next experiment, ROVER was used to combine all possible subsets of our system set, where each individual subset consist of three to nine systems. Five subsets with the lowest combined system WER are listed in table 7. All listed subsets contain systems DA4, LPCC, 15B and DA1, which are the four most complementar systems according to the measure based on *LBWER and DWER matrix* column averaging. The subset with the lowest combined system WER, which consists only of five systems, contains also BSL system, which is the worst system for combination

Table 7: Five best ROVER combinations.

System set	Combined systems								WER [%]
best		DA4		BSL	15B	LPCC	DA1		2.44
2nd best	POW	DA4			15B	LPCC	DA1		2.45
3rd best		DA4		BSL	15B	LPCC	DA1	NOE	2.45
4th best		DA4	30B	BSL	15B	LPCC	DA1	NOE	2.45
5th best		DA4	30B		15B	LPCC	DA1		2.46

according to the measures. We must, however, bring out that proposed measures are correct only to measure suitability of a system for its combination with all other systems. The measures handicap BSL system mainly because of its very low complementarity with systems POW, 30B and ENG as can be seen, for example, in figure 3 (dark fields in BSL row). None of these systems is, however, included in the system subset with the lowest combined system WER. In opposite, brighter fields in BSL row indicates that BSL system is quite complementary to other four systems.

Note that WER of the best ROVER system, which is 2.44%, corresponds to 15.9% relative WER improvement with respect to the best individual system POW and to 5.8% relative WER improvement with respect to ROVER combining all nine systems.

Table 8: Five best ROVER combinations - WER for different conditions.

Condition	**Clean**	**Seen noises**			**Unseen noises**			**Seen**
SNR level	-	**20dB**	**10dB**	**00dB**	**20dB**	**10dB**	**00dB**	**cond.**
POW	1.11	1.70	4.55	48.50	1.50	3.70	37.03	2.90
ROVER 9	1.14	1.41	4.14	49.55	1.35	3.38	37.38	2.59
System set								
best	1.21	1.32	3.85	50.16	1.27	3.27	36.69	2.44
2nd best	1.11	1.27	3.96	49.12	1.26	3.27	36.27	2.45
3rd best	1.21	1.31	3.91	50.66	1.22	3.30	37.54	2.45
4th best	1.14	1.30	3.92	51.20	1.29	3.32	38.50	2.45
5th best	1.18	1.25	3.99	50.34	1.22	3.26	37.44	2.46

Table 8 shows WER of combined systems listed in the table 7 for individual noisy conditions. This table can be compared with table 2 showing WERs for individual systems and for ROVER combining all nine systems. For all five listed combined systems, the highest improvement is observed for seen noises SNR 20dB and 10dB (8/9 of data with respect to which we were searching for the optimal system combination). We also observe good generalization for unseen noises for the same SNR levels.

5 Complementarity measures for set of systems

In the previous section, we have shown some connection between complementarity of recognition systems, their suitability for system combination and LBWER and DWER measures corresponding to these systems. Values from *LBWER and DWER matrices* 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.

Several complementarity measures for a set of systems are proposed in this section and the correlation between proposed measures and actual WER of combined system is shown in experiments.

5.1 Average Lower Bound Word Error Rate (ALBWER)

In the section 4.4, we have expresed the presumption that the smaller values out of the diagonal (and perhaps also on the diagonal) of *LBWER matrix* the better a system set should be for combination. In the previous section, average of LBWER 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 *LBWER matrix* to obtain measure of overall complementarity among systems in a set. Averaging is given by equation:

$$ALBWER(S) = \frac{\sum_{i \in S} \sum_{j \in S} LBWER(i,j)}{|S|^2} \tag{3}$$

where S is a set of systems and $|S|$ denotes number of systems in this set. In this definition, WERs of individual systems (values on the diagonal) are also included in the average. Alternative definition excluding individual WERs from averaging can be expressed by following equation:

$$ALBWER'(S) = \frac{\sum_{i \in S} \sum_{j \in S, j \neq i} LBWER(i,j)}{|S|^2 - |S|} \tag{4}$$

Note that both measures ALBWER and ALBWER' become similar for higher number of elements (systems) in the set as the ratio between number of values in matrix diagonal and values out of diagonal becomes smaller.

5.2 Average Dependent Word Error Rate (ADWER)

This measure is defined in the same manner as ALBWR measure. The only difference is that values from the *DWER matrix* are averaged instead of *LBWER matrix* according to following equation:

$$ADWER(S) = \frac{\sum_{i \in S} \sum_{j \in S} DWER(i,j)}{|S|^2} \tag{5}$$

Again, alternative measure ADWER', where individual WERs are excluded from averaging, is given by equation:

$$ADWER'(S) = \frac{\sum_{i \in S} \sum_{j \in S, j \neq i} DWER(i,j)}{|S|^2 - |S|} \tag{6}$$

5.3 Average Sum of LBWER and DWER (ALBWERDWER)

This measure is a combination of the previous two measures given by equation:

$$ALBWERDWER(S) = \frac{\sum_{i \in S} \sum_{j \in S} LBWER(i,j) + DWER(i,j)}{|S|^2} \quad (7)$$

Every sum of LBWER and DWER in averaging can be regarded as measure of error dependency similar to the LBWER where *dependent errors* are, however, counted twice. This is in agreement with presumption that *dependent errors* are the worse case of *simultaneous errors* (see section 4.4). Again, alternative measure ALBWERDWER' excluding individual WERs from averaging can be defined in the same manner as measure ALBWER' (equation 4).

5.4 Geometric Average of Lower Bound Word Error Rate (GLBWER)

This measure is similar to ALBWER, however, geometric average is used instead of arithmetic average. The measure is defined by following equation:

$$GLBWER(S) = \prod_{i \in S} \prod_{j \in S} LBWER(i,j)^{\frac{1}{|S|^2}} \quad (8)$$

Note that $\frac{LBWER(i,j)}{100}$ can be interpreted as a probability of *simultaneous error* made by systems i and j. Under the assumption that these probabilities are independent for each different pair of systems i and j, GLBWER measure is related to probability that all systems make *simultaneous error* at the same time.

If two particular systems in a system set make no *simultaneous error*, GLBWER measure for the set will be equal to zero. This however does not imply zero WER for combined systems (at least for ROVER combination).

Measures GLBWER', GDWER, GDWER', etc. can be defined in the obvious way. In our experiments, we will show that measures based on geometric average do not differ significantly from those based on arithmetic average for the real data.

5.5 Experimental setup

In the experiments with system set complementarity measures, the same training and testing data described in section 4.5 are used. *Seen conditions test data* are used for estimation of LBWER and DWER matrices. All individual systems again differ only in feature extraction part. Otherwise each system follows the description in section 4.5. Two different sets each consisting of eleven individual systems are used in these experiments to investigate generalization of proposed complementarity measures.

Systems from the first set will be referred as *systems with different features*. These systems are identical to those described in section 4.5, in addition two systems using following new features, which are again derived from BSL features, were included to the system set:

- **W15** - 15ms window is used to compute spectrum of each frame instead of 25ms window. These features allow for more resolution in time in comparison with BSL features

- **W35** - 35ms window is used to compute spectrum of each frame instead of 25ms window. Here feature vector of each frame represents longer time period, however we do not gain more resolution in spectrum (as could be expected), since the spectrum of each frame is smoothed by the 23 band Mel filter bank used also for BSL features.

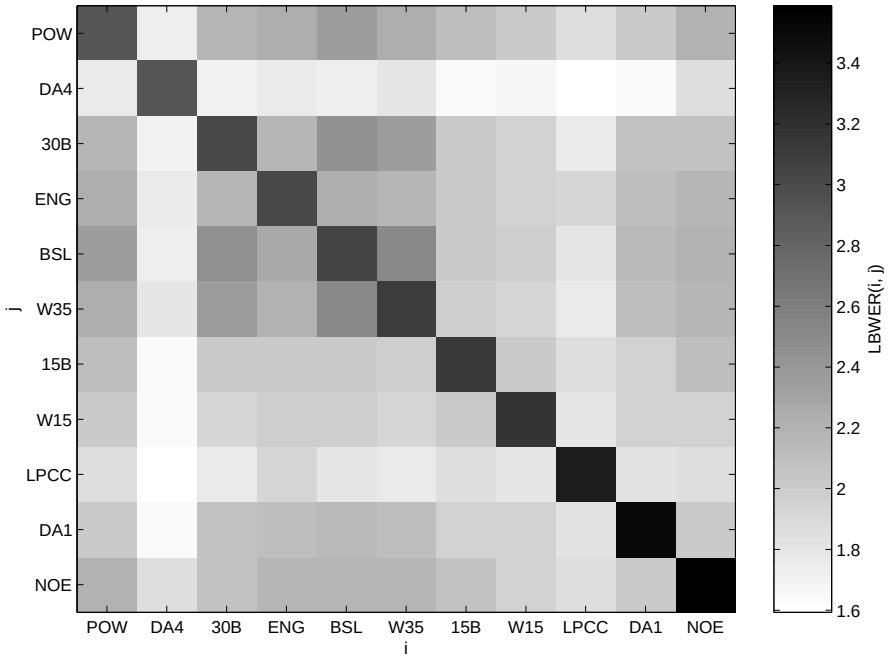


Figure 5: LBWER matrix for systems with different feature.

Graphical representation of *LBWER matrix* for set of *systems with different features* is shown in figure 5. Relatively low LBWER values in column corresponding to system W15 indicates good complementarity of this newly added system with other systems in the set.

Systems from the second set will be referred as *systems with missing bands MFCC*. Features used by these systems are similar to BSL features where, however, log energies of certain bands of Mel filter bank are ignored (always three consecutive bands). Instead of DCT transform, PCA derived on training data is used to decorrelate output of preserved filter bank bands. Features for individual systems differ only in selection of bands that are ignored. Eleven of such systems are used in our experiments:

- **M1-3** - 1st, 2nd and 3rd band of Mel filter bank is ignored
- **M3-5** - 3rd, 4th and 5th band of Mel filter bank is ignored
- **M5-7** - 5th, 6th and 7th band of Mel filter bank is ignored
- ...
- **M21-23** - 21st, 22nd and 23rd band of Mel filter bank is ignored

Graphical representation of LBWER matrix for set of systems with missing bands is shown in figure 6.

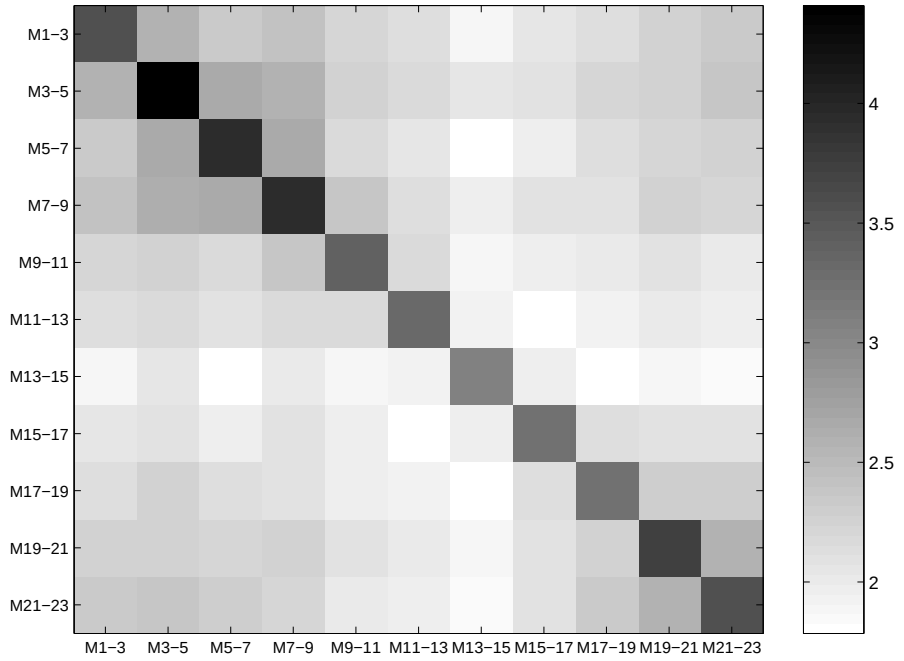


Figure 6: LBWER matrix for systems with missing bands MFCC.

5.6 Correlation between combined system WER and system set complementarity measures

The following experiments were carried out to investigate correlation between proposed system set complementarity measures and corresponding combined system WERs. From the *systems set with different features*, all subsets consisting of three and eight systems were combined using ROVER and corresponding WERs were evaluated. Similarly, all subsets of three and eight systems were combined for the set of *systems with missing bands MFCC*. Combination of three and eight systems was chosen to show how complementarity measures are correlated with combined system WER for combinations of only few (three) and larger number (eight) of systems.

Figure 7a shows WER of combined systems where subsets of *systems with different features* are combined. Each dot corresponds to one combination of three systems and each cross corresponds to one combination of eight systems. Big bold cross represents combination of all eleven systems. Axis Y represents WER of combined system. Combined systems are ordered by their WER on X axis. As can be seen in figure 7a, in average, combinations of eight systems perform better then combinations of three systems, however, the best combinations of three systems (the dots most on the left) perform much better then the worst combinations of eight systems (crosses most on the right). There are few combinations of three systems performing worse than the best individual system POW with WER 2.90% and in

opposite the best such cobmination performes almost as well as the combination of all eleven systems. Of course, the most interesting part of the figure is on the left from the big bold cross, where systems are outperforming the combination of all eleven systems. The best combined system in the figure with WER 2.38 is one of the combinations of eight systems.

Similar figure 7b shows WER of combined systems where subsets of *systems with missing bands MFCC* are combined. When combining *missing bands MFCC* systems, the goal is to outperform BSL system with WER 3.04, which uses information from all the bands. As can be seen in the figure, combination of all eleven systems with WER 2.67 reaches the goal. Many combinations of three systems perform worse than BSL system, on the other hand, there are combinations of three systems outperforming even the combination of all eleven systems. The best combined system in the figure with WER 2.51 is one of the combinations of eight systems.

Figure 8 shows correlation between WER of combined system (X axis) and average of WERs of corresponding individual systems (axis Y). Again, each dot, cross and big cross in the figure corresponds to one combination of three, eight and eleven systems respectively. Figure 8a shows that for *systems with different features* no significant correlation can be observed. Therefore, we can conclude that for this system set, WERs of individual systems are not important for selection of systems suitable for combination. In figure 8b, for *systems with missing bands MFCC*, some correlation between combined system WER and average WER of individual systems can be seen. [5] As was shown in figure 6, for this system set, systems with lower WER were generally more suitable for combination, however, it does not mean that average WER of individual systems is the good measure of system complementarity. We will see that the proposed complementarity measures are much more correlated with corresponding WER of combined system.

In the following experiments, we will see a correlation between proposed system set complementarity measures and corresponding combined system WER. We will compare different measures and make conclusions about their performances for combinations of small number and larger number of systems. Properties of the measures are again demonstrated on combinations of three and eight systems from set of *systems with different features* and set of *systems with missing bands MFCC*. Presented results of these experiments may not seem to be sufficient to make some of the following conclusions, however, trends similar to those presented here were observed also for different number of combined systems and for different sets of systems.

Figure 9 shows the correlation between combined system WER and corresponding *Average Lower Bound Word Error Rate (ALBWER)* measure computed according to equation 3. For both system sets and for combinations of three and eight systems, visible correlation is observed between ALBWER measure and combined system WER. For systems with missing band MFCC, much higher correlation is observed in comparison to that seen in figure 8b.

Figure 10 shows the correlation between combined system WER and ALBWER' measure (alternative definition of ALBWER measure excluding WERs of individual systems from averaging), which is computed according to equation 4. In comparison to ALBWER, this measure is less correlated with combined system WER for combinations of three systems (the dots are more spread around the line on which they would ideally lay). This could be, however, specific only to ROVER combination

[5] Note that we must look at combinations of three systems and eight systems separately.

with majority voting used in our experiments, where voting based on decision of only few systems can be unreliable and actual WER of individual systems can by more important. In opposite, comparing figures 9a and 10a, WER of combinations of eight systems seems to be more correlated with ALBWER' measure than with ALBWER measure.

It can be seen in figure 10 that dots representing combinations of three systems and crosses representing combinations of eight systems are concentrated around two separate lines. Therefore, values of ALBWER' 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 ALBWER' measure to choose which systems will be good for combination. The same rule applies for all other proposed complementarity measure.

Figures 11 and 12 show complementarity measures based on averaging of values of *DWER matrix* according to equations 5 and 6 respectively. Again, we observed that ADWER measure is more correlated with three systems combination WERs than ADWER' and, in opposite, eight systems combination is more correlated with ADWER' measure. An interesting finding is that for higher number of combined systems, measures ADWER and ADWER' show higher correlation with WER of combined system than measures ALBWER and ALBWER'.

In section 5.3, we proposed measure ALBWERDWER averaging sums of corresponding values from *LBWER and DWER matrix*. However, experiments with this measure did not show any particular advantage of using this measure. Results obtained for this measure look simply as a compromise between ALBWER and ADWER' measure.

In section 5.4, measures based on geometric average of values from *LBWER and DWER matrix* were proposed. Figure 13 demonstrate results obtained in experiment with GLBWER measure given by equation 8. Again, no particular advantage of using geometric average was observed. Results obtained for these measures were almost identical to those obtained for corresponding measures based on arithmetic average, specially for higher number of combined systems[6].

6 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 the 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 measures of recognition systems complementarity, which are based on measurement of error dependency of individual system outputs. First, methods for measuring complementarity of two systems were proposed. These

[6]Comparing figures 13 and 9, relative positions of crosses are almost identical.

measures 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 these measures is used as an extension allowing to measure the complementarity of a system subset. Correlation between these measures and actual performances of combined systems was shown in experiments, which indicates that these measures can be advantageously used to select systems suitable for their combination.

References

[1] F. Jelinek. "Statistical Methods for Speech Recognition", MIT Press, 1998.

[2] B. Gold and N. Morgan. "Speech and Audio Signal Processing", New York, 1999.

[3] 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.

[4] R.G. Leonard. "A database for speaker-independent digit recognition", Proc. ICASSP'84, pp. 42.11.1-4.

[5] H. G. Hirsch, D. Pearce. "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRW ASR2000, "Automatic Speech Recognition: Challenges for the Next Millennium", Paris, France, September 2000.

[6] M. J. Hunt. *A statistical approach to metrics for word and syllable recognition* J. Acoust Soc. Am., vol. 66(S1), S35(A), 1979

[7] H. Hermansky. "Perceptual linear predictive (PLP) analysis of speech", Journal of the Acoust. Soc. of Am., vol. 87, pp. 1738-1752, 1990.

[8] P. Jain, H. Hermansky and B. Kingsbury, "Distributed Speech Recognition Using Noise-Robust MFCC And TRAPSEstimated Manner Features" , Proc. of ICSLP 2002, Denver, Colorado, September 2002. Table 6: Aurora 3 Relative improvement for QIO.

[9] F. Grézl, L. Burget, P. Jain, J. Černocký, "Improving TRAPS features using LDA", in *12th International Czech and Slovak Scientific Conference (RADIOELEKTRONIKA'02)*, Bratislava, Slovak Republic, Bratislava, May 15-16, 2001

[10] R. Duda, P. Hart, "Pattern Classification and Scene Analysis", New York: John Wiley & Sons, 1973

[11] K. Fukunaga. "Introduction to statistical pattern recognition". Academic press, Inc., Boston, USA, 2 edition, 1990.

[12] H. Bourlard and N. Morgan. "Connectionist Speech Recognition - A Hybrid Approach". Kluwer Academic Press, 1994.

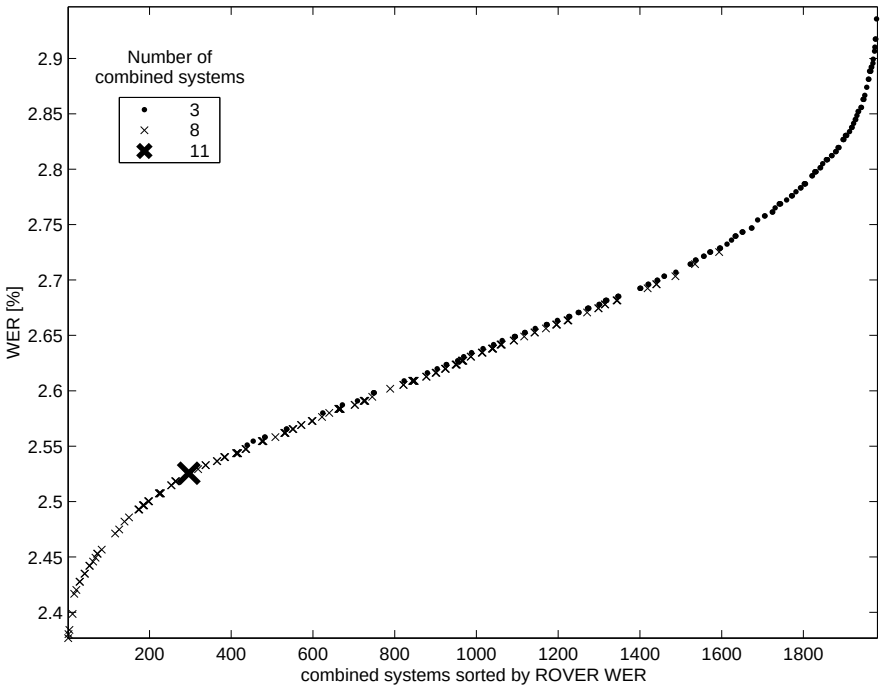
[13] H. Hermansky,D. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," Proc. ICASSP, Istanbul, June 2000.

[14] D.W.P. Ellis, R. Singh and S. Sivadas, Tandem Acoustic Modeling in Large-Vocabulary Recognition , in Proc. ICASSP 01, Salt Lake, City, Utah, USA, May 2001.

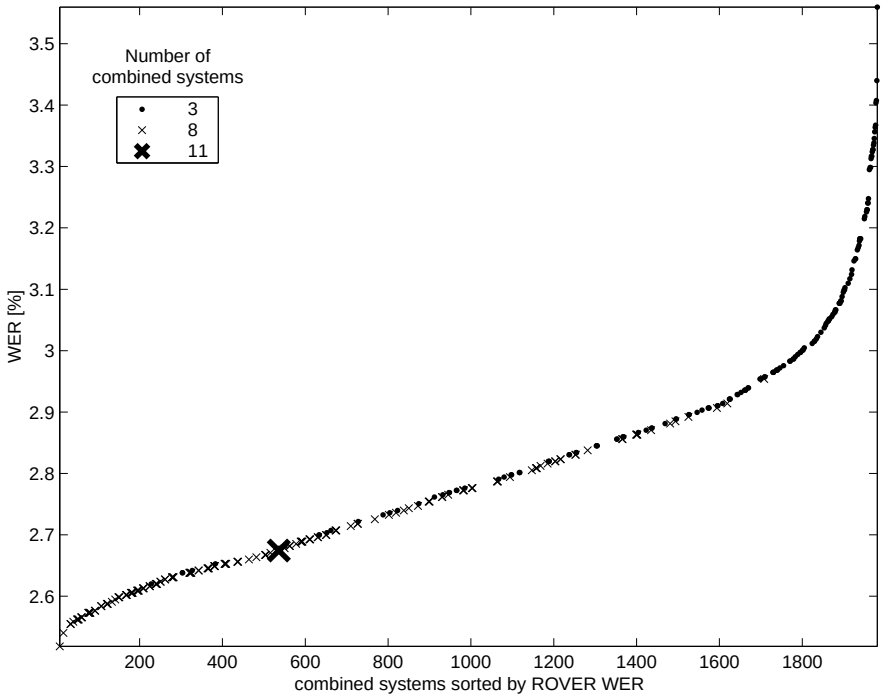
[15] Y. Normandin, "Maximum Mutual Information Estimation of Hidden Markov Models", Automatic Speech and Speaker Recognition", pp. 57-81, 1996.

[16] W. Chou, C. H. Lee, and B. H. Juang, "Minimum Error Rate Training Based on N-Best String Models", in Proceedings of the 1993.

[17] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)", in Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 1997, pp. 347–354.

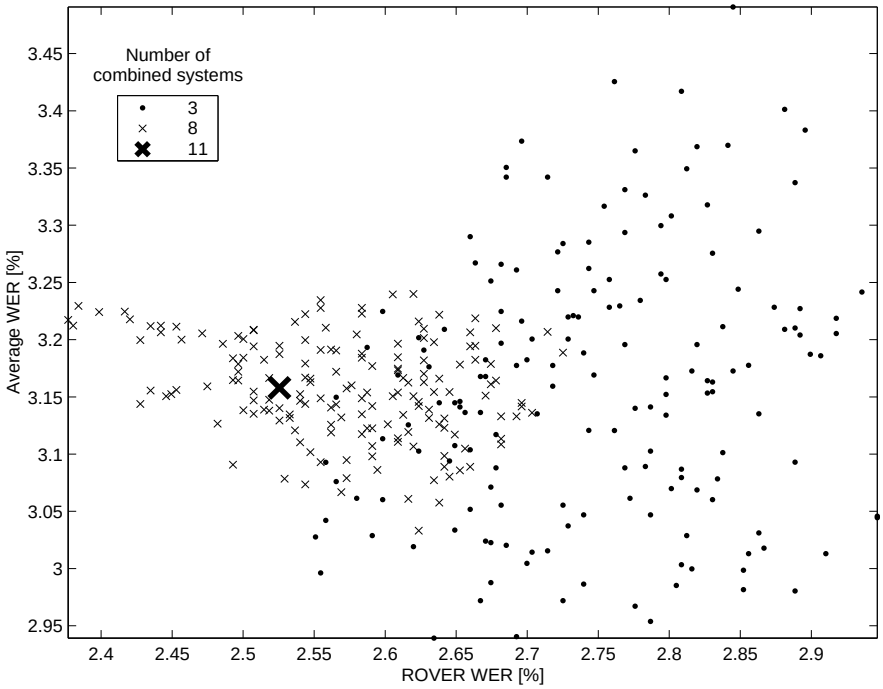


a) Different features

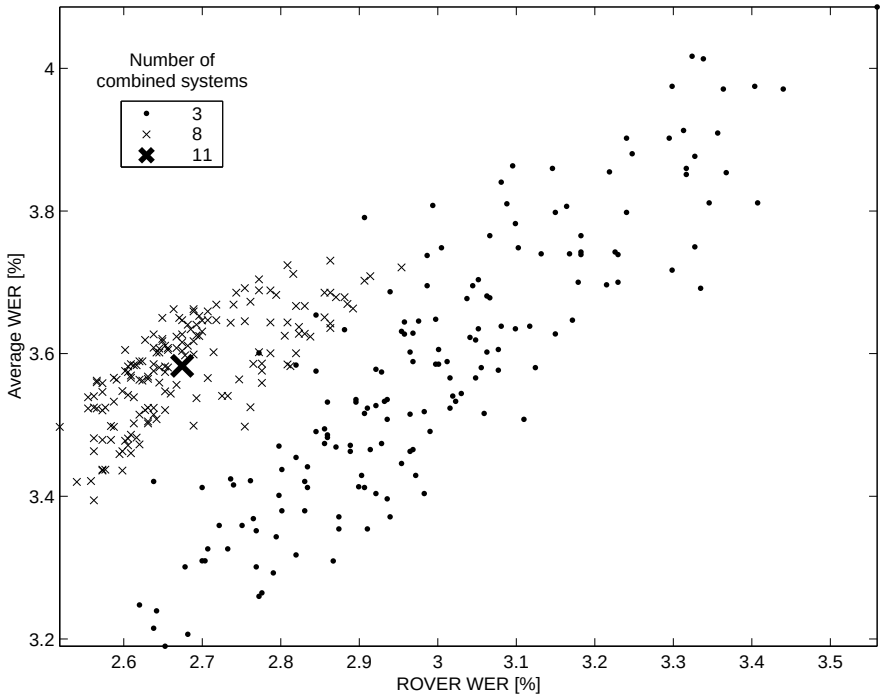


b) Missing band MFCC

Figure 7: ROVER WER for combinations of 4 and 7 ans all 11 systems.

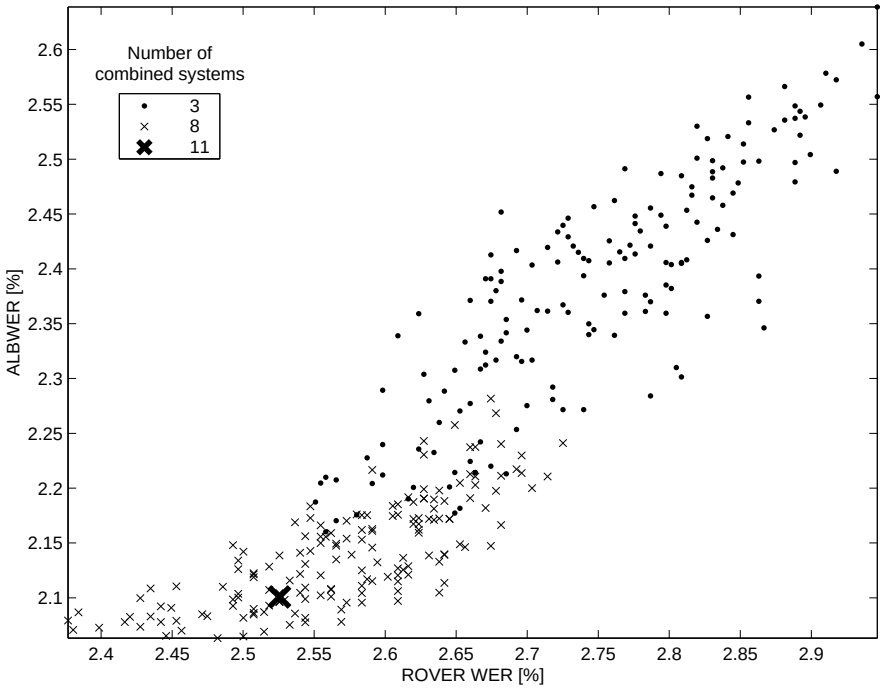


a) Different features

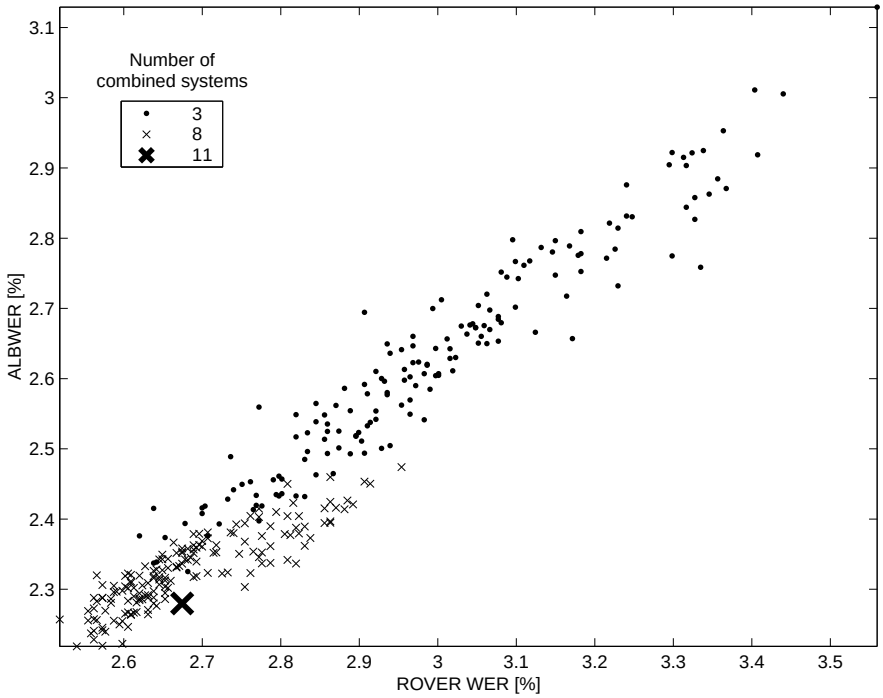


b) Missing band MFCC

Figure 8: Correlation between average WER and ROVER WER.

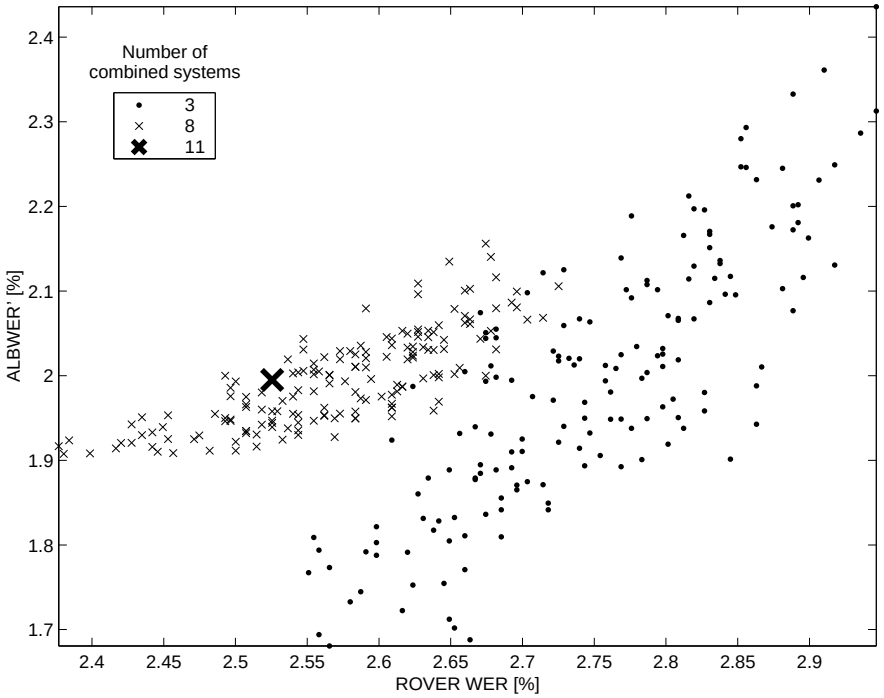


a) Different features

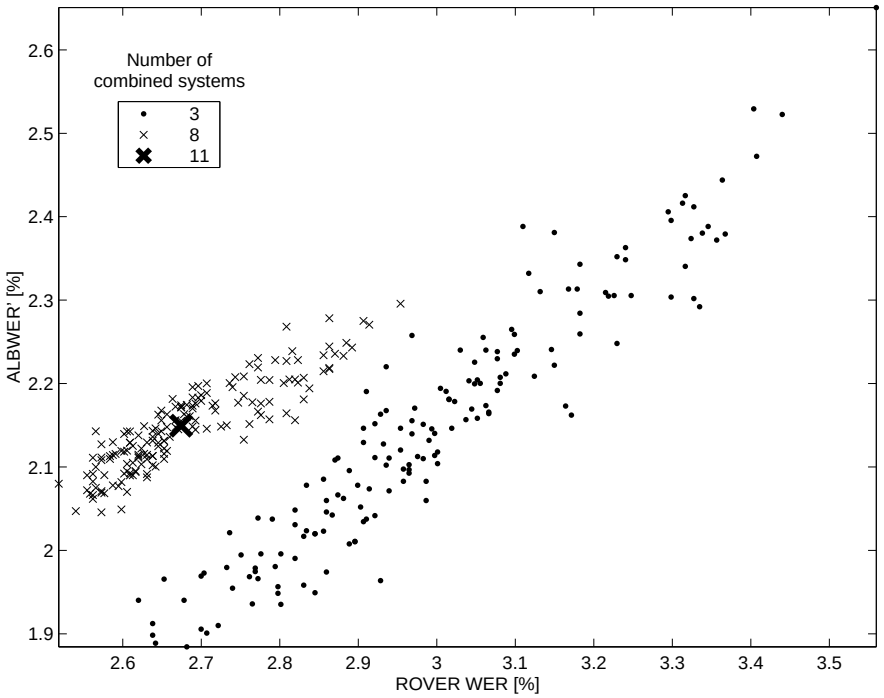


b) Missing band MFCC

Figure 9: Correlation between ALBWER and ROVER WER.

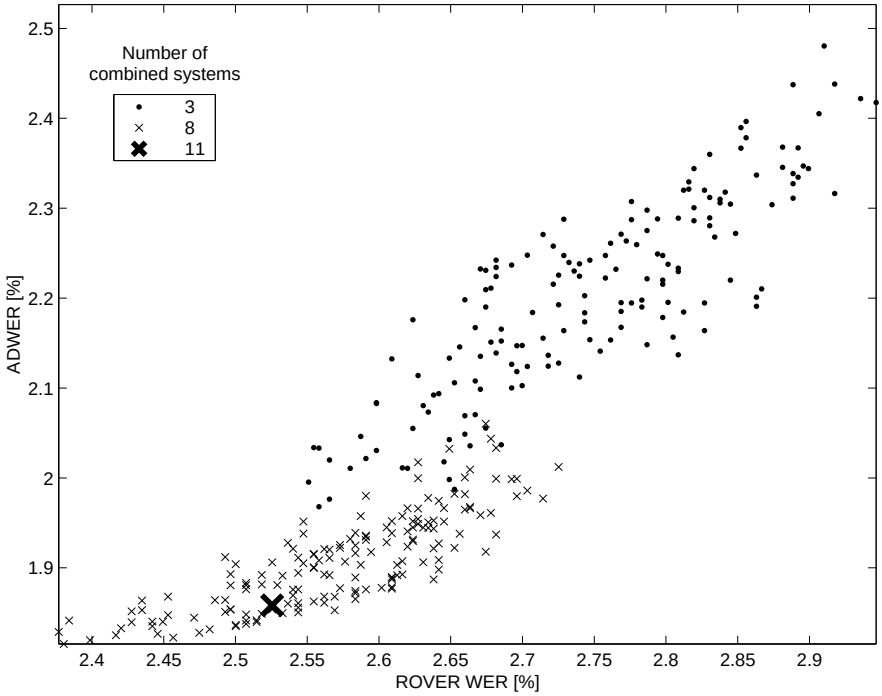


a) Different features

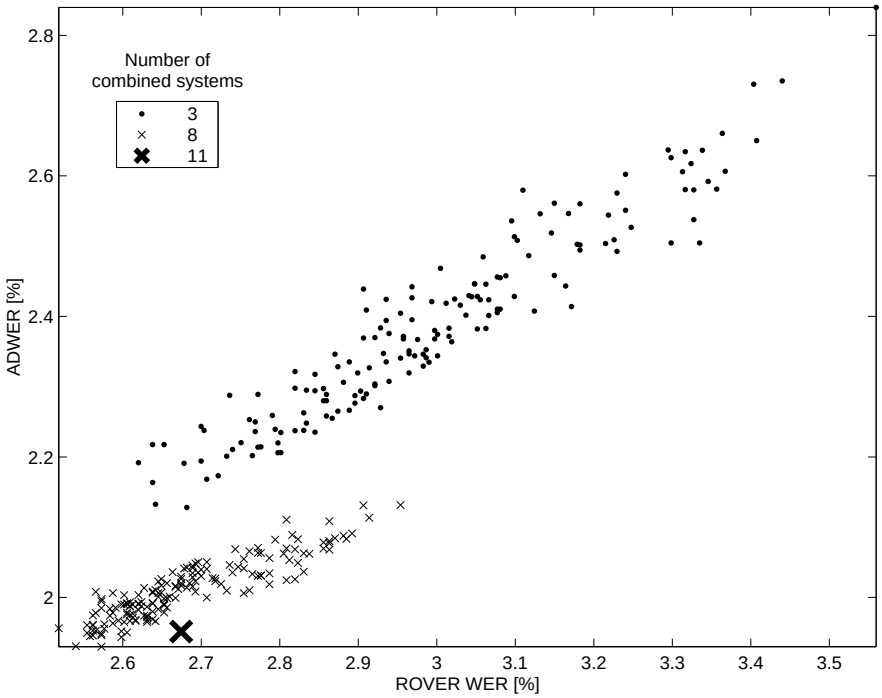


b) Missing band MFCC

Figure 10: Correlation between ALBWER' and ROVER WER.

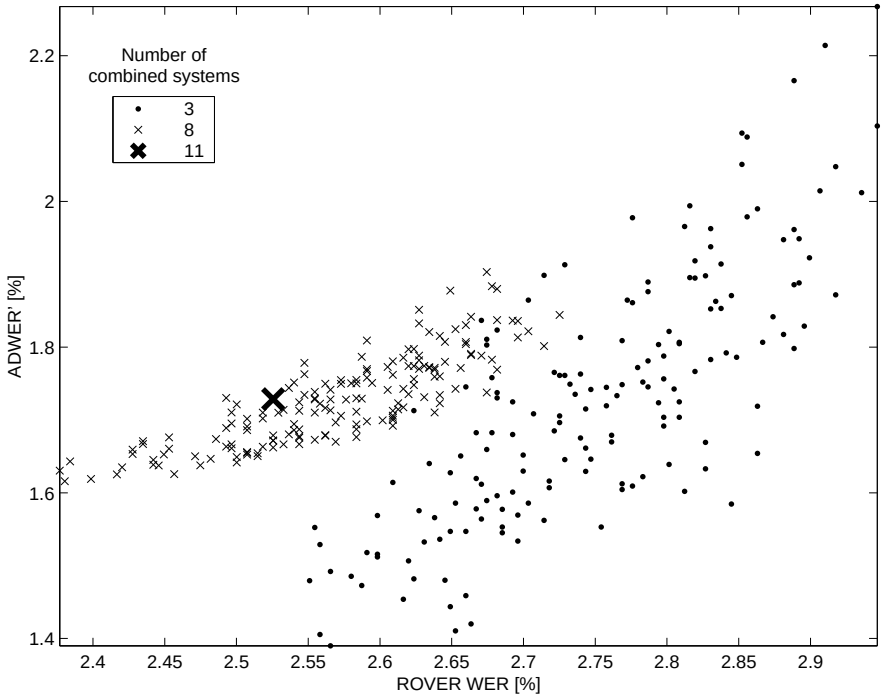


a) Different features

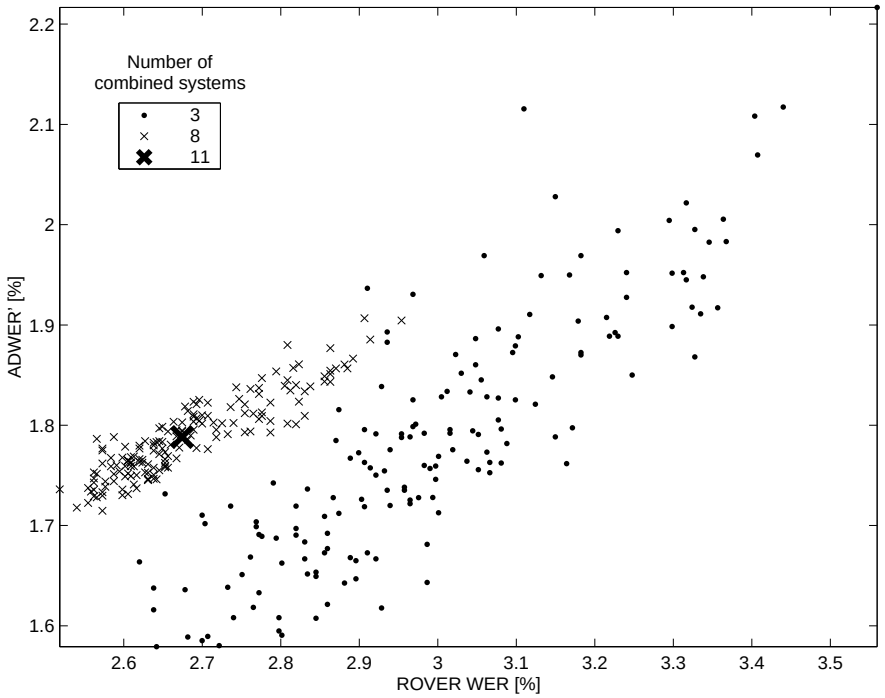


b) Missing band MFCC

Figure 11: Correlation between ADWER and ROVER WER.

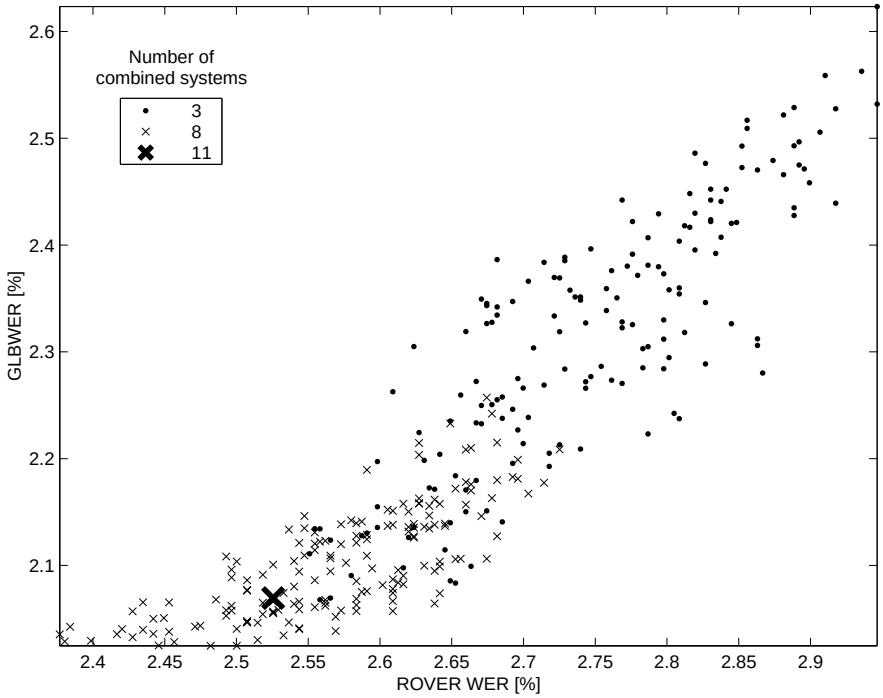


a) Different features

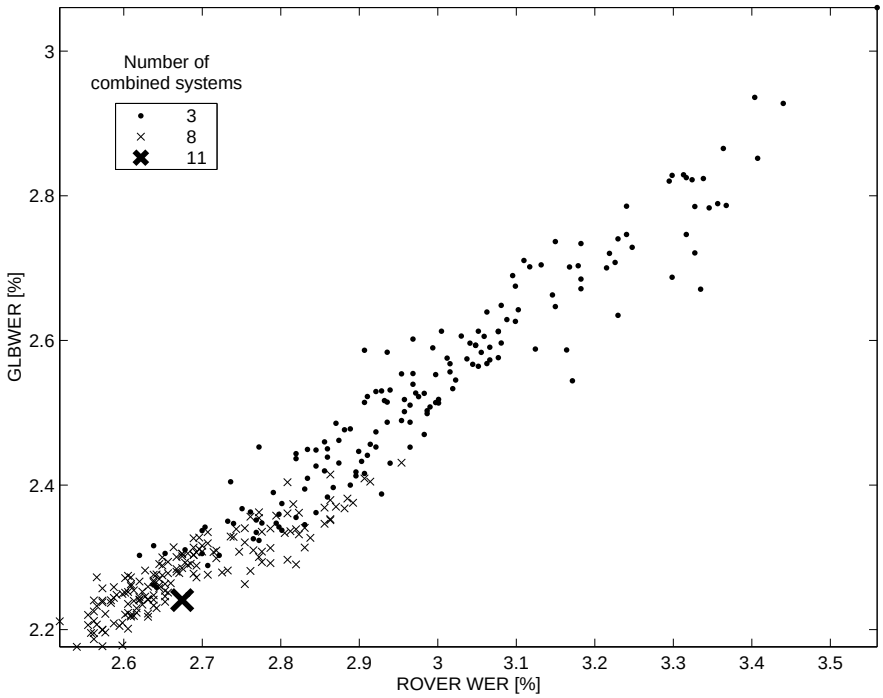


b) Missing band MFCC

Figure 12: Correlation between ADWER' and ROVER WER.



a) Different features



b) Missing band MFCC

Figure 13: Correlation between GLBWER and ROVER WER.