Elektrotehniški vestnik 75(4): 229–234, 2008
Electrotechnical Review, Ljubljana, Slovenija

Criteria for the Evaluation of Automated Speech-Recognition Scoring Algorithms

Simon Dobrišek

University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, 1000 Ljubljana, Slovenia
E-mail: simon.dobrisek@fe.uni-lj.si

Abstract. Variations of the basic string-alignment algorithm are commonly used for the detection and classification of speech-recognition errors. In this procedure, reference and system-output hypothesis speech transcriptions are first aligned using the string-alignment algorithm that is based on primitive edit operations. The edit operations needed to transform one transcription into the other are then tallied as speech-recognition errors. The algorithms normally detect approximately the same total number of errors; however, they can produce different error classifications. This paper investigates these differences and several criterion functions that can be used for the comparison and evaluation of the algorithms for the detection and classification of speech-recognition errors. The proposed criterion functions were used for the experimental evaluation of the standard algorithms that are implemented as part of the CUED HTK and the NIST SCTK, and were used for the detection and classification of phone-recognition errors from the TIMIT speech database.

Key words: speech recognition, scoring algorithm, error classification, classification agreement, confusion matrix

Kriteriji za ovrednotenje samodejnih postopkov ocenjevanja razpoznavalnikov govora

Povzetek. Za odkrivanje in razvrščanje napak pri samodejnem razpoznavanju govora se uporablja več različic osnovnega postopka poravnave nizov simbolov. Pri tem postopku se najprej poravna referenčne in samodejno tvorjene govorne prepise in nato se ugotavlja razlike med njimi v smislu napak vrivanja, izbrisa in zamenjave govornih enot. Omenjene različice ponavadi odkrijejo približno isto skupno število napak, vendar se med sabo razlikujejo po končni razvrstitvi odkritih napak. V članku raziskujemo te razlike in tudi kriterijske funkcije, ki bi jih bilo mogoče uporabiti za primerjavo in ocenjevanje postopkov za odkrivanje in razvrščanje napak pri samodejnem razpoznavanju govora. S predlaganimi kriterijskimi funkcijami smo ocenili razlike med tovrstnimi postopki, ki so del programskih paketov CUED HTK in NIST SCTK ter so bili uporabljeni za odkrivanje in razvrščanje napak pri samodejnem razpoznavanju glasov iz govorne zbirke TIMIT.

Ključne besede: razpoznavanje govora, postopek ocenjevanja, razvrščanje napak, ujemanje razvrstitev, matrika razvrstitev

1 Introduction

The evaluation of automatic speech-recognition (ASR) systems relies on the automated scoring procedures that detect and classify ASR errors by comparing the reference and system-output hypothesis transcriptions of test utterances. The comparison is performed using the string-alignment algorithm that computes the string-edit distances [14, 8, 1, 7] between pairs of speech transcriptions. The algorithm produces the optimum alignment between a pair of speech transcriptions and determines the minimum-cost sequence of the basic edit operations needed to transform one transcription into the other. The basic edit operations are then tallied as ASR errors in terms of speech-unit substitutions, deletions and insertions [4].

The optimum alignment is usually assigned so as to minimize the total ASR-error rate (TER), which is the sum of insertion, deletion and substitution errors, divided by the number of speech units in the reference transcriptions. The minimum number of substitutions, deletions and insertions needed to transform one speech transcription into the other is known as the Levenshtein distance [9] between the two transcriptions. A well-known feature of this distance is that multiple different optimum alignments may exist for the same pair of transcriptions with the same distance value [7]. Consequently, multiple different ASR errors may be detected and classified from the same pair of transcriptions. This reduces the diagnostic value of examining the resulting ASR-error classifications, and it can even hinder the understanding of ASR failure mechanisms.

A variation of the basic string-alignment algorithm that computes the so-called weighted string-edit distance [11] is often used to improve the ASR-error classi-

Received 1 September 2008
Accepted 26 September 2008

fication. This distance model assumes that the basic edit operations are weighted by some edit costs that are symbol dependent. These edit costs can be assigned according to some prior knowledge about the ASR errors. For instance, the edit costs can be assigned according to the a-priori probability of the individual ASR errors, or even according to the time distance between the speech units in a pair of speech transcriptions [2].

In any case, different variations of the string-alignment algorithm produce different resulting ASR-error classifications, and determining which of them are more correct is not a trivial task [2]. Usually, it is not feasible to manually examine all the ASR errors and to make a decision about which of them are detected and classified correctly and which are not. A system for the automated evaluation of hypothesis ASR-error classifications would require a database of the reference ASR-error classifications that are produced from the same pairs of speech transcriptions as the hypothesis ones. To the best of our knowledge, there is no such database that could be used for this purpose.

We decided to investigate whether any indication of which scoring algorithm produces more correct ASR-error classifications can be drawn directly from the classifications themselves. We examined several criterion functions that are well-known in statistics and are commonly used for measuring the agreements between different classifications or clusterings of the same set of objects. We conducted several experiments using different scoring algorithms that produced different ASR-error classifications from the same pairs of speech transcriptions and the obtained classifications were then evaluated using the proposed criterion functions.

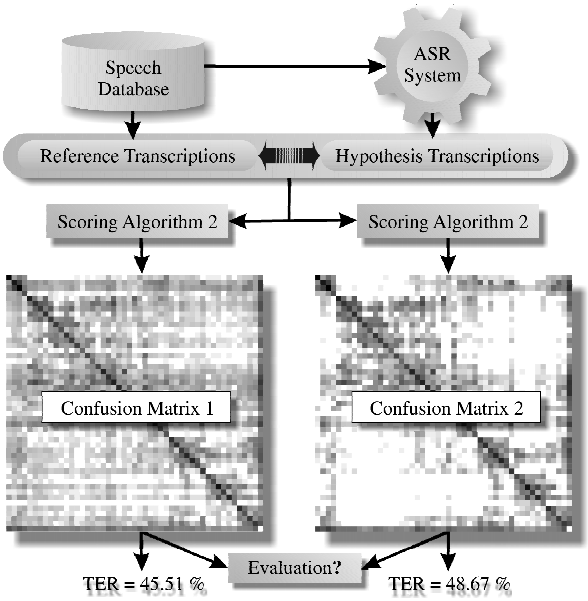


Figure 1. An illustration of the problem investigated in this paper. The two graphically represented confusion matrices were obtained using two different scoring algorithms from the same pairs of phonetic transcriptions of the TIMIT speech recordings. Their different levels of gray correspond to the frequencies of the confusion pairs. The columns and rows of the two matrices correspond to the 48 TIMIT phonetic units. These phonetic units are ordered by their classification into the phone classes (vowels, sonorants, plosives, etc). The bottom row and the right–most column of both matrices correspond to the null symbol, i.e., phone insertions and deletions.

2 Background

Conventional ASR systems rely on subword acoustic models, such as the phonetic hidden Markov models (HMMs), where each HMM represents a context-dependent allophone. Such models are often developed and evaluated separately using the phone-recognition error rate as a performance measure. The phone-recognition error rate is obtained by aligning the reference and hypothesis phonetic transcriptions of the test utterances and tallying the phone-recognition errors in the same manner as discussed above. The scoring algorithms used for the classification of phone-recognition errors are normally the same as those used for the classification of word-recognition errors.

We limited ourselves to an analysis of the phone-recognition-error classifications produced by different scoring algorithms. The reason for this decision was that certain assumptions about phone-recognition errors can undoubtedly be derived from general psycho-acoustic knowledge about the human system of speech production and perception, and these assumptions are the basis for the evaluation of the proposed criterion functions that are more general and can also be used for the evaluation of the word-recognition-error classifications.

The usual ASR-scoring algorithms normally provide not only the total ASR-error rate but also a detailed list of ASR-error classifications, where the individual errors are classified as speech-unit substitutions, insertions or deletions. Each ASR error can be represented as a confusion pair, where the speech-unit deletions are considered to be simply substitutions of speech units with a null unit, and vice verse for insertions. The frequencies of all the confusion pairs are often presented in a matrix form that is called a confusion matrix.

Fig. 1 illustrates the investigated problem. Two different scoring algorithms, which are actually implemented as part of the NIST SCTK [12], calculated similar total phone-recognition error rates from the same pairs of phonetic transcriptions, however, they produced two very different confusion matrices. The two matrices represent two different phone-recognition error classifications and there are several reasons why we would like to evaluate which of them is more correct. One of the reasons is that the more correct error classification has a higher diagnostic value and can thus help in improving the developed

ASR system. Another reason would be that certain ASR error-correction algorithms can improve their ability to correct errors if these are classified more correctly.

3 Statistical analysis of confusion matrices

Many different statistical measures were proposed to measure the degree of association between two categorial variables and most of them are derived from the contingency table [5]. These measures include the chi-squared, phi-squared and G-test, Cramer's V and the lambda statistics, joint entropy, mutual information, etc.

Our confusion matrices are special examples of contingency tables, where the two corresponding variables range over the same set of categories (i.e., speech units). One variable is associated with the reference speech transcriptions and the other with the hypothesis ones. The null symbol, which is used for the representation of deletion and insertion errors, is also considered as a category by itself in the same way as all the other speech units. Measuring the association between two such variables can be interpreted as measuring the agreement between two (speech) classifications, and such an agreement is generally assessed using statistical measures like the raw agreement, the chance-corrected agreement κ, the chance-independent agreement ϕ, etc.

We studied many of the mentioned measures and later in the paper we present some of those that seem to be the most appropriate for the analysis of the ASR scoring results.

Let us focus now on the basic features of the ASR-related confusion matrices. Suppose that we obtained different total phone-recognition-error rates and also different confusion matrices using two different ASR scoring algorithms from the same pairs of reference and hypothesis phonetic transcriptions, as illustrated in Fig.1. From the illustration it is clear that in the left-hand matrix the phone-recognitions errors are distributed more uniformly (randomly) over the pairs of phonetic units than is the case with the right-hand matrix. Considering the nature of the system of human speech production and perception, one would intuitively expect that the less random distribution of phone-recognition errors should be the more correct. This intuition is based on the general psychoacoustic observation that confusions between the phonetic units within the same broad-phonetic classes are more likely to happen than confusions between the phonetic units from different broad-phonetic classes, e.g., it is more likely that a sonorant is confused with another sonorant than with a non-sonorant[3].

On the other hand, the first scoring algorithm a detected considerably lower total number of errors (TER = 45.51%) than the second one (TER = 48.67%). However, only from these two values we cannot make any obvious assumption about which of the two algorithms produces the more correct error classification. More precisely, a smaller number of detected errors does not necessarily indicate that the corresponding algorithm detects errors more correctly.

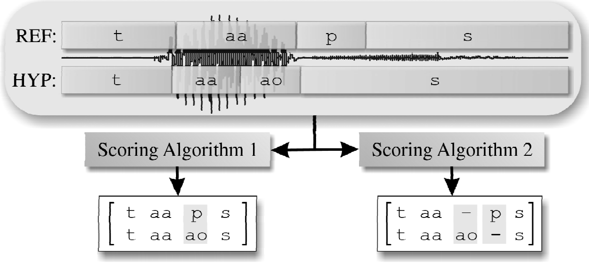


Figure 2. Two ASR scoring algorithms produced two different alignments of the reference and hypothesis phonetic transcriptions. The alignments are represented as the sequences of the vertically aligned confusion pairs. The symbol '$-$' denotes the null symbol that is used for the representation of deletion and insertion errors. The highlighted confusion pairs represent the phone-recognition errors that were detected and classified by the two ASR scoring algorithms.

Let us illustrate this with an example. Fig 2 shows the two phonetic alignments of the reference and hypothesis phonetic transcriptions that were produced by the two ASR scoring algorithms from Fig.1. The first algorithm detected fewer errors than the second one; however, if the two phonetic segmentations are considered, the second confusion pairs are obviously more correct than the first ones. In other words, it is obvious from the two phonetic segmentations that the vowel `[ao]` is not substituted with the plosive `[p]`; it is inserted and then the plosive is deleted.

4 Measuring the agreement between two classifications

The total ASR error rate is the only nominal measure that we have for now, and it is derived directly from the confusion matrix. This measure is normally used for the evaluation of ASR systems, and, as it is clear from the above example, it cannot be used for the evaluation of the ASR scoring algorithms that are used for the evaluation of ASR systems. The total error rate depends only on the sum of all the detected errors; it does not consider their distribution, which is arguably important. We studied many other nominal measures that can also be derived from the confusion matrix and consider the distribution of errors. These measures are well known in statistics and are used for measurements of the agreement between two partitions (classifications/clusterings) of a set of objects [6, 10].

The first group of measures is derived directly from the confusion matrix. Among many such measures we focused on the following:

- the chance-corrected agreement κ,

- Cramer's V measure of association,
- Goodman-Kruskal's λ,
- the normalized mutual information NMI,
- the likelihood ratio G-test.

The Cohen's κ coefficient is a statistical measure of inter-classifier agreement. If the two classifiers are in complete agreement then $\kappa = 1$. If there is no agreement between the two classifiers then $\kappa \leq 0$. Cramer's V is a measure of association that is based on adjusting the Pearson χ^2 significance to factor out sample size. It ranges from 0 to 1, where 1 indicates a strong relationship between variables and 0 indicates no relationship. The Goodman-Kruskal's λ is a measure of the proportional reduction in error. The values for λ range from zero (no association between two variables) to one (perfect association). The normalized mutual information is used for the test of independence. If one variable is completely dependent on the other, then the NMI takes its maximum value of 1. If the two variables are independent, then the NMI equals 0. The likelihood ratio G-test is a test for goodness of fit to a distribution and for independence in contingency tables: the higher the value the more dependent are the two variables. When one variable is unrelated (i.e., only randomly related) to the other variable the likelihood ratio equals 0.

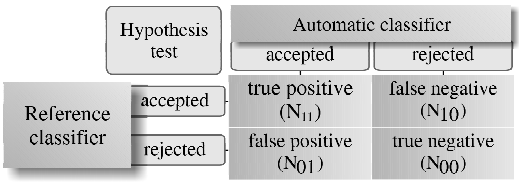
Among the above measures only κ takes its maximum value of 1 if and only if the total error rate equals zero. In this case the confusion matrix is diagonal. The other three measures take their maximum value of 1 when the confusion matrix has exactly one non-zero element in each row and column. This means that they can also take the value of 1 when the total error rate is not zero and some speech units are always confused with the same, other speech units.

The next group of measures is normally used for measuring the agreement between two observers that make statistical decisions on a given hypothesis [10]. One of them is considered as a reference observer and the other can make Type I and Type II errors, also known as false-positive and false-negative decisions. The results of their decisions are recorded and analysed using 2×2 contingency tables. For our evaluation problem, such a table is defined as shown in Fig. 3.

We examined two hypotheses that one could formulate in the context of the discussed evaluation problem. The first hypothesis is used for the strict comparison of two speech classifications, i.e., the measure of agreement between two classification takes its maximum value if and only if the total number of errors is zero. The hypothesis is formulated as follows:

$H_1^{(a)}$: *The given speech unit is classified in the selected speech class.*

The decisions on the above hypothesis are assumed to be taken individually for each of the considered speech classes and the results are summed over all of them. Such a hypothesis is, for instance, used for the statistical analysis of confusion matrices that is implemented as part of the LingPipe suite of Java libraries for the linguistic analysis of human language.



Hypothesis test		Automatic classifier	
		accepted	rejected
Reference classifier	accepted	true positive (N11)	false negative (N10)
	rejected	false positive (N01)	true negative (N00)

Figure 3. The 2×2 contingency table that is defined with the second group of statistical measures.

The number of true/false positive/negative decisions on the above hypothesis can be derived directly from the usual confusion matrix. Let $\mathbf{C}$ denote a $k \times k$ confusion matrix, whose elements m_{ij} denote the number of detected confusions between the i-th and j-th speech classes. Let then $n = \sum_{i,j} m_{ij}$ denote the total number of all the classified speech units, including the number of all the null units that are associated with insertion and deletion errors. The number of all the four possible decisions on $H_1^{(a)}$ are then defined as follows:

$$\begin{gathered} N_{11} = \textstyle\sum_i m_{ii} \ , \quad N_{10} = N_{01} = n - N_{11} \ , \\ N_{00} = k\,n - (N_{11} + N_{10} + N_{01}) \end{gathered} \tag{1}$$

The second hypothesis is normally used for the comparison of two clusterings of the same set of objects. The hypothesis is formulated as follows:

$H_1^{(b)}$: *The given two speech units are classified in the same speech class.*

Let $n_{i\cdot}$ and $n_{\cdot j}$ be the i-th row and j-th column sums of $\mathbf{C}$, respectively. The number of the four possible decisions $H_1^{(b)}$ can also be derived directly from the confusion matrix and they are then defined as follows:

$$\begin{gathered} N_{11} = \textstyle\sum_{i,j} \binom{m_{ij}}{2} \ , \\ N_{10} = \textstyle\sum_i \binom{n_{i\cdot}}{2} - N_{11} \ , \quad N_{01} = \textstyle\sum_j \binom{n_{\cdot j}}{2} - N_{11} \ , \\ N_{00} = \binom{n}{2} - (N_{11} + N_{10} + N_{01}) \end{gathered} \tag{2}$$

Many different measures of agreement between two classifications/clusterings are defined as functions of the number of all the four possible decisions [5, 6]. We focused on the following most widely used:

- the Fowlkes-Mallows's index,
- the Jaccard index,
- the Adjusted Rand index,
- the Yules Q and Yules Y indexes.

All the above indexes take their maximum value of 1 when the number of false-positive and false-negative decisions both equal 0, i.e., the two classifications are in complete agreement. On the other hand, if the number of true-positive and true-negative decisions both equal 0, then the values of these indexes are equal to or less than zero, i.e., the two classifications are in complete disagreement.

We computed and compared all the presented statistical measures from the confusion matrices obtained by the different ASR scoring algorithms from the same pairs of phonetic speech transcriptions. Our basic assumption here is that these measures should indicate which ASR scoring algorithm produces the more correct ASR-error classification; the higher the value, the less randomly the ASR errors are distributed over the confusion pairs and thus, the more correct is the ASR-error classification.

Besides the above statistical measures and the usual total error rate, TER, we also observed several simple ratios that provide additional information about the distribution of the classified ASR errors. These ratios are the following:

- the broad-class error rate (BCER),
- the Levenshtein error ratio (LER),
- the indels error ratio (IDER),

The broad-class error rate, BCER, is derived from the usual TER, where the ASR errors that occur within the broad speech classes are considered as ASR hits. The BCER is thus always lower or equal to the TER. According to the mentioned psychoacoustic observations, a lower value of the BCER may indicate more correct ASR-error classifications. The Levenshtein error ratio, LER, is simply the relative ratio between the TER obtained by a given ASR scoring algorithm and the TER obtained by the ASR scoring algorithm that is based on the computation of the Levenshtein distance. The LER thus provides information on how many more ASR errors than the minimum possible number of detected ASR errors the given ASR scoring algorithm detected. The last ratio, IDER, is the sum of all the detected insertion and deletion errors divided by the sum of all the detected errors.

We cannot make any obvious assumptions about the expected values of the LER and IDER; however, one would naturally expect that the ASR scoring algorithm should not detect too many additional ASR errors and that the majority of detected errors are substitutions and not perhaps insertions and deletions. This means that low values are preferred for the LER and IDER.

5 ASR scoring algorithms

The ASR scoring algorithm that is based on the computation of the Levenshtein distance between speech transcriptions detects the minimum possible number of errors that can, in general, be detected from given pairs of speech transcriptions. Let us denote this basic variant of the scoring algorithm as LSED.

The ASR scoring algorithms implemented in some widely used ASR scoring packages, like the NIST SCTK [12] and CUED HTK HResults [15], are based on the string-edit distance that assigns a cost value to substitutions that is higher than the cost value assigned to insertions and deletions. The original motivation for using such costs is in the redistribution of the classified ASR errors from substitutions, on one hand, to insertions and deletions on the other. This redistribution increases the so-called ASR correctness, which is defined as the percentage of all the speech units in the reference transcriptions classified as ASR hits.

The basic NIST SCTK scoring algorithm assigns a cost value of 4 to substitutions and 3 to insertions and deletions. Similarly, the CUED HTK HResults assigns 10 to substitutions and 7 to insertions and deletions. Let the SCTK algorithm be denoted as NSED and the HTK one as HSED.

The last investigated ASR scoring algorithm is also implemented as part of the NIST SCTK and is based on the so-called time-mediated string-alignment algorithm. This algorithm is the only one that also considers the time-of-occurrence of individual speech units. In this algorithm, the costs that are assigned to the basic string-edit operations are time dependent and are based on the beginning and ending times of the speech units. Let this algorithm be denoted as NTMA. A more detailed explanation of this algorithm is given in the documentation that is part of this toolkit

6 Experimental results

For the experiments we built a simple phone recognizer using the CUED HTK toolkit [15] and the TIMIT speech database [13]. The recognizer was deliberately designed to be simple and to produce a relatively high error rate. This is because our goal was not to improve the ASR accuracy, but to study the differences in the ASR-error classifications obtained by different ASR scoring algorithms.

The structure of the phone recognizer is very conventional. The left-to-right three-state HMMs with mixtures of eight Gaussian densities per state were used for all 48 context-independent monophone models. No language model was used. The parameters of the monophone HMMs were estimated from the training set of the TIMIT database using the Baum-Welch algorithm, as proposed in [15]. Hypothesis phonetic transcriptions and segmentations were then generated for all the speech recordings in the TIMIT database using the usual Viterbi beam-search algorithm without pruning.

All the confusion matrices that were obtained using the four ASR scoring algorithms mentioned above were then analysed using all the considered statistical measures and ratios.

The obtained results are given in Table 1. The val-

		ASR scoring algorithm			
		LSED	HSED	NSED	NTMA
I	κ	0.548	0.552	**0.553**	0.523
	CV	0.564	**0.579**	0.578	0.560
	λ	0.530	0.537	**0.538**	0.510
	NMI	0.500	0.519	0.518	**0.541**
	GT	360533	378426	376703	**394888**
IIa	FM	0.562	0.566	**0.567**	0.540
	J	0.391	0.395	**0.396**	0.369
	AR	0.553	0.557	**0.558**	0.530
	YQ	0.985	0.986	**0.986**	0.983
	YY	0.844	0.845	**0.846**	0.833
IIb	FM	**0.411**	0.403	0.407	0.384
	J	**0.256**	0.253	0.256	0.237
	AR	**0.390**	0.382	0.386	0.361
	YQ	**0.942**	0.937	0.939	0.929
	YY	**0.705**	0.695	0.698	0.679
	TER	**45.51**	45.70	45.52	48.67
	BCER	28.03	28.76	28.35	**25.20**
	LER	-	0.42	**0.02**	6.94
	IDSR	**29.28**	36.23	34.77	34.42

Table 1. The values of all the considered evaluation measures (given in abbreviations) obtained by different ASR scoring algorithms.

ues in bold are those that are considered to be the best for a given measure. As discussed earlier, intuitively, the LSED algorithm should produce the least, and NTMA the most, correct ASR-error classifications. However, as it is clear from the table, only three measures, i.e., the NMI, BCER and the G-test, are considered to be the best for the NTMA, and the LSED obtained the best values in the IIb group of statistical measures. We attributed this to the fact that the NTMA detected considerably more ASR errors (LER $= 6.94\%$) than any other algorithm, and this may be reflected in most of the proposed statistical measures, especially in the group IIb.

From this perspective it seems that the most relevant evaluation measures are those in the first group. Their values indicate that the HSED, NSED, NTMA produce more correct ASR-error classifications than the basic LSED algorithm. However, these results also indicate that the NTMA is perhaps not an optimal ASR scoring algorithm and that some further improvements are possible.

7 Conclusions and future work

Several statistical measures were investigated that may be used for the development and evaluation of ASR scoring algorithms. The presented results of the evaluation of the four widely-used ASR scoring algorithms indicate that further improvements are possible in this research area. The presented measures provide an evaluation tool that can help with such improvements. We believe that any improvements will be indicated at least by an increase in the values of the NMI, BCER and the G-test and a decrease in the value of the LER.

In our future work we will focus on the development of an improved ASR-scoring algorithm that is based on the string-edit costs that are time and symbol dependent, and we will use the presented measures for its evaluation.

8 References

[1] A. V. Aho, "Algorithms for finding patterns in strings", *Handbook of Theoretical Computer Science*, J. Leeuwen (ed.), Elsevier Science Publishers, pp. 255–300, 1990.

[2] G. Doddington, "Word alignment issues in ASR scoring" *Proceedings of ASRU*, US Virgin Islands, pp. 630–633, 2003.

[3] A. Cutler, A. Weber, R. Smits and N. Cooper, "Patterns of English phoneme confusions by native and non-native listeners", *Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3668–3678, 2004.

[4] D. Gibbon, R. Moore, and R. Winski (eds.), *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, Walter de Gruyter Publishers, Berlin, 1997.

[5] L. A. Goodman, and W. H. Kruskal, *Measures of association for cross classifications.* Springer-Verlag, New York, 1979.

[6] L. Hubert, and P. Arabie, "Comparing partitions", *Journal of Classification*, vol. 2, pp. 193–218, 1985.

[7] J. B. Kruskal and D. Sankoff, "An Antology of Algorithms and Concepts for Sequence Comparisons", *Time Warps, String Edits and Macromolecules: The Theory and practice of Sequence Comparison*, D. Sankoff and J. Kruskal (eds.), CSLI Publications, pp. 256–310, 1999.

[8] W. Masek and M. Paterson, "A faster algorithm computing string edit distances", *Journal of Computer System Sciences*, vol. 20, no. 1, pp. 18–31, 1980.

[9] V. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals", *Soviet Physics–Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[10] T. Hill and P. Lewicki, *STATISTICS Methods and Applications*, StatSoft, Tulsa, OK, 2007.

[11] T. Okuda, E. Tanaka and T. Kasai, "A method of correction of garbled words based on the Levenshtein metric", *IEEE Transactions on Computers*, vol. 25, no. 2, pp. 172–177, 1976.

[12] *The NIST Speech Recognition Scoring Toolkit (SCTK) Version 2.3.rc2*, [Web page], `http://www.nist.gov/speech/tools`, The NIST Speech Group. National Institute of Standards and Technology (NIST), USA. [Accessed June 6th, 2008].

[13] *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, (TIMIT), [CDROM], National Institute of Standards and Technology (NIST), USA, 1990.

[14] R. A. Wagner and M. J. Fisher, "The string-to-string correction problem", *Journal of the Association for Computing Machinery*, vol. 21, no. 1., pp. 168–173, 1974.

[15] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, L. Xunying, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4).* Cambridge University Engineering Department, Cambridge, UK, 2006.

Simon Dobrišek is a Research Fellow and Teaching Assistant in the Laboratory of Artificial Perception, Systems and Cybernetics at the Faculty of Electrical Engineering in Ljubljana.