Predicting the Acoustic Confusability between Words for a Speech Recognition System using Levenshtein Distance

A. Zgank, Z. Kacic
Institute of Electronics and Telecommunications, University of Maribor,
Smetanova ul. 17, SI-2000 Maribor, Slovenia, phone:+386 2 220 7206
andrej.zgank@uni-mb.si

***Abstract*—This paper proposes a new method for calculating acoustic confusability between words for automatic speech recognition. Acoustic confusability is one of the key elements influencing speech recognition accuracy. The proposed method is based on Levenshtein distance, calculated on phonetic transcriptions from the speech recognizer's vocabulary. The method was evaluated in an indirect way. The experiments were carried out on four different sets of context-dependent acoustic models. The proposed method successfully predicted the acoustic confusability between words from the speech recognizer's vocabulary.**

***Index Terms*—Acoustic modeling, automatic speech recognition, human computer interaction, Levenshtein distance.**

I. INTRODUCTION

An automatic speech recognition system is one of the prerequisite modules in case when developing a system for supporting natural human-computer interaction [1]. Examples of such human-computer interaction are Interactive Voice Response (IVR) systems, virtual avatars, intelligent ambient systems, etc.

A speech recognizer's vocabulary must be created during the process of developing a spoken dialog for HCI. The acoustic similarity between words within the vocabulary results in acoustic confusability, which decreases the speech recognition accuracy. In order to control and reduce this effect, this paper proposes a method for predicting the acoustic confusability of words for speech recognition. This method is based on Levenshtein distance [2], [3] calculated from phonetic transcriptions of words taken from speech recognizers' vocabulary. The proposed method is especially useful when new words are added to the vocabulary during a system's development phase, as they can be modified in those cases of worse acoustic confusability. In such a way, the quality of service (e.g. IVR), remains at the same level. Another usage possibility is to apply the proposed method for the prediction, which set of acoustic models (i.e. different type or complexity) [4], [5] would produce the best speech recognition results for a given test set/vocabulary [6]. The defined method could be also applied as similarity metric for cross-lingual speech recognition [7].

The proposed method for predicting the acoustic confusability of words is language independent. These experiments were carried out with the Slovenian language, but this method could be applied to any other language with non-trivial grapheme to phoneme conversion, resulting in a phoneme set with various members. All experiments were carried out on isolated words, as the usage of a test scenario with a statistical language model could influence the evaluation of the proposed method.

Manuscript received March 18, 2012; accepted June 3, 2012.

This work was partially supported by Slovenian Research Agency under contract number P2-0069 "Advanced methods of interaction in telecommunication".

II. ACOUSTIC CONFUSABILITY AND LEVENSHTEIN DISTANCE

Acoustic confusability between words is one of the key elements influencing the performance of a speech recognition system [3]. One of the possibilities for estimating the acoustic confusability of a new word is to calculate the acoustic similarity of words based on acoustic models. The drawback of such a method is that it can be complex and, in addition, needs full access to the parameters of the acoustic models. This is not usually the case for commercial speech recognition systems. The solution is to use a metric based on a speech recognizer's vocabulary, which is usually accessible.

The proposed predictive method originates from Levenshtein distance (LD) [3], calculated on phonetic transcriptions of words within the speech recognizer's vocabulary. The Levenshtein distance gives the number of operations needed to transform one phonetic transcription into another one. The available transformation operations are the insertion, deletion or substitution of a phoneme within the transcription. For example, the Levenshtein distance between the phonetic transcription of the English words "house" (/h o u s e/) and "houses" (/h o u s e s/) is 1, since you need only one insertion (last /s/) to transform the phonetic transcription of "house" to the transcription of "houses". Such a Levenshtein distance is dependent on the length of the phonetic transcription, thus the normalization of distance is used for the length of the phonetic transcription. The normalized Levenshtein distance (NLD) is defined as

$$NLD_i = \frac{LD(w_i, w_j)}{l_{\max}}, \quad (1)$$

where $LD(w_i, w_j)$ denotes the Levenshtein distance between words i and j and $l_{\max}$ denotes the number of phonemes from the longer word. The resulting NLD takes values between 0 and 1. Previously detailed analysis of speech recognition errors have shown that the largest part of misrecognitions occurred between two or three words in the set, usually those that are acoustically similar. This was the starting-point for the proposed method. Thus the proposed acoustic confusability (*AC)* of word *i* is defined as

$$AC_i = \alpha \cdot \min(NLD_i) + \beta \cdot \overline{NLD}, \quad (2)$$

where NLD_i denotes the normalized Levenshtein distance, and α and β denote the empirically defined weights. They were set to values 0.9 and 0.1 for the speech recognition systems involved in the experiments.

When a new word should be added to a test scenario, the acoustic confusability is calculated for the phonetic transcriptions of all those particular words included within the speech recognizer's vocabulary.

III. SPEECH DATABASE

The speech recognition systems involved in the experiments were developed using the Slovenian 1000 FDB SpeechDat(II) database. The databases from the SpeechDat family are applied for constructing various voice-driven telecommunication services and cover, at the moment, more than 50 different languages. The Slovenian SpeechDat(II) database includes recordings of 1000 speakers over fixed telephone lines. For each speaker 43 different utterances were recorded. The structure of the speakers in the database is demographically balanced. The training set consists of 800 speakers and the test set of 200. Various test scenarios with isolated or connected words can be used for evaluating the speech recognition system. The most frequently used test scenarios are:

- voice-mail command words,
- isolated and connected digits,
- yes/no answers,
- city names,
- phonetically balanced words.

The size of the vocabulary for these test scenarios varies between 2 and 1491.

One of key factors that influence on the acoustic confusability of words, and thus the complexity of a speech recognition system, is the number of phonemes included in the acoustic models. The Slovenian SpeechDat(II) database has 46 different phonemes. If less frequent phonemes are mapped into similar more frequent ones, using the acoustic-phonetic knowledge, the number of phonemes can be reduced. One set with 39 phonemes and one set with 25 phonemes were trained in such a way. This modeling approach influences the acoustic confusability between different words and results in an increased number of training examples per particular phoneme.

In addition to these acoustic models based on phonemes, a separate set was trained based on 25 Slovenian graphemes. The advantage of grapheme acoustic models is that they don't include any additional errors introduced during grapheme to phoneme conversion, which can be very difficult for some languages.

An evaluation of proposed method was carried out on a voice-mail command words test scenario containing 31 different words with 1070 recordings. This test scenario was the most suitable one in the speech database, as it contained the highest number of recordings per isolated word. Four test cases were prepared. Three words (approx. 10%) were randomly excluded from the vocabulary for the construction of the first test case. This presented the evaluation baseline. The excluded 3 words (W1, W2 and W3) where then added to the baseline vocabulary one at a time and the acoustic confusability was calculated for these cases. The evaluation was done in an indirect way, using speech recognition results.

IV. EXPERIMENTAL SETUP

The speech recognition systems involved in the experimental setup were based on monolingual COST 278 MASPER scripts [8]. In such a way, an identical training procedure was used for all the different sets of acoustic models.

The feature extraction was based on 12 mel-cepstral coefficients and energy. In addition to the basic 13 features, delta and delta-delta features were also included. The final feature vector had 39 elements. The window size was 25 ms, and the window shift 10 ms. Cepstral mean normalization was applied to improve the front-end robustness.

Three state left-right continuous density Hidden Markov Models topology was used for the acoustic modeling. Three step approach was applied for training the acoustic models. The initial set contained context-independent acoustic models with one Gaussian probability density function per state. The forced realignment procedure was carried out using these models, with the goal of improving the speech transcriptions included in the acoustic training. On the typical training set, less than 0.1 % of utterances were excluded using the forced realignment procedure.

The improved transcriptions were included in the second step, where the acoustic models were trained from scratch using the model specific initial values. After the initialization, the number of Gaussian mixtures per state was increased to 32 in a stepwise manner. The resulting acoustic models were again involved in the forced realignment procedure. Less than 0.06 % of utterances were excluded from the training set this time.

The context-dependent (triphone) acoustic models were developed during the last step of acoustic modeling. The phonetic decision-tree based clustering was applied, to reduce the number of free acoustic models' parameters, which should be estimated during the training. The phonetic decision trees were induced with the broad phonetic classes generated from a data-driven approach based on the phoneme similarity estimation [9]. The number of Gaussian mixtures per state in the triphone acoustic models was

incrementally increased to 16. This was the final version of the acoustic models used for the evaluation.

The complete acoustic modeling training procedure was carried out for 4 different sets of acoustic models. Context-dependent acoustic models were trained with 25, 39 and 45 phonemes, respectively and grapheme based context-dependent acoustic models with 25 elements. The comparable complexity (i.e. number of free acoustic models' parameters) for all the final context-dependent acoustic models was controlled with the threshold parameter during the decision-tree based clustering procedure.

V. RESULTS

The evaluation of the experimental setup for predicting acoustic confusability between words during automatic speech recognition was carried out in two steps. First, the acoustic confusability of words W1, W2, and W3 was calculated for the given speech recognizer's vocabulary for all 4 sets of acoustic models. The results are presented in Table I.

TABLE I. ACOUSTIC CONFUSABILITY OF WORDS W1, W2, AND W3 FOR DIFFERENT SETS OF ACOUSTIC MODELS.

Acoustic models	**Acoustic confusability**		
	W1	**W2**	**W3**
GR-25	0.582	0.239	0.489
PH-25	0.583	0.239	0.487
PH-39	0.587	0.239	0.612
PH-45	0.587	0.239	0.612

The results for the calculated acoustic confusability predicted that the highest number of misrecognitions would occur for word W2. Its acoustic confusability was 0.239 for all four different sets of acoustic models. This level of acoustic confusability was the result of word W2, which differed with the next most similar word by only one insertion. The acoustic confusability for word W3 varied between 0.487 for PH-25 acoustic models and 0.612 for the PH-39 and PH-45 acoustic models, respectively. The word W1 had acoustic confusability of 0.582 for GR-25 acoustic models and 0.587 for the PH-39 and PH-45 acoustic models. This range of values for word W1 predicted the lowest number of misrecognitions for this word. The pairwise identical values of acoustic confusability for the PH-39 and PH-45 acoustic models confirmed, that the additional 6 phonemes in the PH-45 set are infrequently found in vocabularies. Such acoustic models are more difficult to train. The predicted acoustic confusability between words showed insignificant distinction between the grapheme (GP-25) and phoneme (PH-25) types of acoustic models with comparable models' complexities.

In the second step, the previously calculated acoustic confusability between the words was indirectly evaluated with the speech recognition results for different acoustic models on a given test scenario using the baseline and three new words in the vocabulary. The speech recognition results are given as word error rates (WER), which is defined as

$$WER(\%) = \frac{E}{N} \cdot 100, \qquad (3)$$

where E denotes the number of misrecognized words in the test set and N denotes the number of all words in the test set. The speech recognition results are presented in Table II.

TABLE II. WORD ERROR RATE FOR BASELINE SET AND THREE NEW WORDS WITH DIFFERENT SETS OF ACOUSTIC MODELS.

Acoustic models	**WER(%)**			
	Baseline	**W1**	**W2**	**W3**
GR-25	1.78	1.70	2.63	2.12
PH-25	1.99	1.90	2.93	2.32
PH-39	2.09	2.00	3.03	2.32
PH-45	2.20	2.10	3.13	2.42
Average	*2.02*	*1.93*	*2.93*	*2.30*

The baseline test scenario achieved word error rates within the range from 1.78 % (GR-25) to 2.20% (PH-45), with an average WER of 2.02 % over all four sets of acoustic models. These baseline speech recognition results are comparable with other similar experimental systems [8], [10]. When the new word W1 was added to the test scenario, the WER slightly decreased for all four setups although the number of words in the vocabulary increased. The relative improvement of WER was between 4.50% and 4.76%. The acoustic confusability of W1 already indicated that this word acoustically-phonetically differed from other words in the test scenario. The additional factor that probably led to this performance improvement was the length (9 phonemes/graphemes) of the word W1, which was above the average.

The word error rate increased statistically significantly for word W2. The best WER for W2 was 2.63 % (GR-25) and the worst was 3.13 % (PH-45). The relative difference of WER to the baseline was between 42.27% and 47.75%. A detailed analysis of the speech recognition results showed, that the misrecognitions of word W2 in the majority of cases occurred with the most acoustically similar word from the vocabulary. The results of indirect evaluation confirmed the predicted acoustic confusability regarding word W2.

The test scenarios with word W3 achieved WER between 2.12 % (GR-25) and 2.42 % (PH-45), which represents a small degradation of the speech recognition performance in comparison with the baseline results. These results are in correlation with the predicted acoustic confusability regarding word W3.

A secondary result of this second step of evaluation was the comparison between different sets of acoustic models. The grapheme type of acoustic models consequently proved to have better performances than the phoneme type with the same complexity. This confirms that non-trivial grapheme to phoneme conversion introduces an additional level of errors to the speech recognition performance. The increased number of phonemes (from 25 to 45) decreased the performance of the experimental setup. The probable cause for this degradation was the low frequency of the added phonemes, which made it difficult to correctly estimate the acoustic models' parameters. The resulting acoustic models lost one part of the generalization effect as a consequence.

VI. CONCLUSIONS

This paper presented a novel approach for predicting the

acoustic confusability between words within a given test scenario for a speech recognition system. Such an approach can be used as support during the designing of a spoken dialog system, with the goal of improving the speech recognition results in advance. The evaluation showed that the proposed metric of acoustic confusability between the words, successfully predicted the speech recognition performance.

Our future work will be oriented towards including additional knowledge about acoustic-phonetic characteristics to an acoustic confusability metric.

REFERENCES

[1] J. Juhar, A. Cizmar, M. Rusko, M. Trnka, G. Rozinaj, R. Jarina, "Voice operated information system in Slovak", *Compuiting and Informatics*, Slovak Acad. Sciences Inst. Informatics, vol. 26, no. 6, pp. 577– 603, 2007.

[2] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals", *Soviet Physics Doklady*, no. 10, pp. 707–717, 1966.

[3] B. Ziółko, J. Gałka, T. Jadczyk, D. Skurzok, "Modified weighted Levenshtein distance in automatic speech recognition", in *Proc. of 16th Conf. on Applications of Mathematics in Biology and Medicine*, Krynica, pp. 116–120, 2010.

[4] A. Žgank, "Three-Stage Framework for Unsupervised Acoustic Modeling Using Untranscribed Spoken Content", *ETRI Journal*, vol. 32, no. 5, pp. 810–818, 2010. [Online]. Available: http://dx.doi.org/10.4218/etrij.10.1510.0092

[5] K. Vicsi, G. Szaszak, "Using prosody to improve automatic speech recognition", *Speech Communication*, Elsevier, vol. 52, no. 5, p. 413–426, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2010.01.003

[6] N. Theera-Umpon, S. Chansareewittaya, S. Auephanwiriyakul, "Phoneme and tonal accent recognition for Thai speech", *Expert Systems with Applications*, Pergamon-Elsevier Science, vol. 38, no. 10, pp. 13254–13259, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2011.04.142

[7] P. Kasparaitis, "Lithuanian Speech Recognition Using the English Recognizer", *Informatica*, Inst. Mathematics & Informatics, vol. 19, no. 4, pp. 505–516, 2008.

[8] A. Žgank, Z. Kačič, F. Diehl, K. Vicsi, G. Szaszak, J. Juhar, S. Lihan, "The COST 278 MASPER initiative: Crosslingual speech recognition with large telephone databases", in *Proc. of LREC 2004*, Lisbon, 2004.

[9] A. Žgank, B. Horvat, Z. Kačič, "Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity", *Speech Communication*, – Elsevier, vol. 47, no. 3, pp. 379–393, 2005.

[10] B. Dropuljic, D. Petrinovic, "Development of Acoustic Model for Croatian Language Using HTK", *Automatika*, Korema, vol. 51, no. 1, pp. 79–88, 2010.