Phonetic Distance Measures for Speech Recognition Vocabulary and Grammar Optimization

Michael Pucher[1], *Andreas Türk*[2], *Jitendra Ajmera*[3], *Natalie Fecher*[3]

[1]Telecommunications Research Center Vienna, Vienna Austria
[2]Speech and Signal Processing Lab, TU Graz, Graz, Austria
[3]Deutsche Telekom Laboratories, Berlin, Germany

pucher@ftw.at, tuerk@ftw.at, Jitendra.Ajmera@telekom.de, Natalie.Fecher@telekom.de

Abstract

This paper reports on the correlation between word confusion matrices from Word-Error-Rate (WER) experiments and different phonetic distance measures. The investigated phonetic distance measures are based on the minimum-edit-distances between phonetic transcriptions and the distances between Hidden-Markov-Models (HMM). We show that phonetic distance measures are correlated with word confusion. The correlations between word confusion of a speech recognizer and phonetic distance are useful for a speech recognition grammar developer or a spoken dialog system designer in developing efficient grammars and dialogs. Furthermore the measures can be used for evaluating the quality of grammars in terms of phonetic confusability of words/utterances or interpretations. An extension of these measures to grammar optimization is discussed.

Index Terms: Speech recognition, speech communication

1. Introduction

Many different phonetic distance/similarity measures have been proposed. Here we are interested in measures that can be defined on resources that are available on a speech recognition platform and do not need further data. This includes phonetic dictionaries and acoustic models. Therefore we decided to investigate models that operate on these two types of resources.

To investigate phonetic distance models for phonetic dictionaries three distance measures based on the minimum-edit-distance [1] are used. The first measure is the standard minimum-edit-distance. The second measure uses adjusted substitution weights, based on phoneme-feature-overlap [2]. The third measures uses adjusted substitution weights, based on a phonetic similarity measure that is derived through perceptual similarity tests [3].

Various phonetic distance measures that make use of acoustic HMM models have been proposed [4, 5, 6, 7]. Here we use a distance that finds the minimum total path between the states of two HMMs, where the local distance between two states is given by the Kullback-Leibler (KL) divergence of the two states.

To know which type of measure should be used by a speech recognition grammar designer to reduce WER by reducing phonetic confusability, we optimize the phonetic minimum-edit-distance measures on a number recognition task. This data is the development test set for the phonetic minimum-edit-distance measures and the test set for the HMM and minimum-edit-distance measures. The HMM-based measure and the minimum-edit-distance measure, which are not optimized serve thereby as an upper and lower bound for the performance of the measures.

Phoneme	Features
s	Consonant, Obstruent, Fricative, Continuant, Anterior, Strident, Coronal, *Unvoiced, Fortis*
z	Consonant, Obstruent, Fricative, Continuant, Anterior, Strident, Coronal, *Voiced, Lenis*

Table 1: *Features of 's' and 'z'.*

In the evaluation the optimized phonetic minimum-edit-distance measures are correlated to word confusion and confusions of interpretations of a speech recognizer for a spoken dialog system. This system is implemented on a Nuance dialog platform. It provides information on fixed and mobile telephone and internet tariffs, and allows internet problems to be reported. It generates log-files and records the recognition result in each dialog state.

The confusion test data for the evaluation was collected within a usability test of a prototype implementation for a prequalifying application of the Deutsche Telekom. This usability test was performed in cooperation with Siemens AG, Corporate Technology, Competence Center "User Interface Design". The recognized utterances are extracted from logging data of the usability test. Reference transcriptions are made on the basis of the speech data that is logged and processed by the recognizer.

Interpretations are natural language interpretations that the system assigns to a certain recognized utterance. The data from the spoken dialog system is the test set for the phonetic minimum-edit-distance measures. We show that word confusions and word distances are correlated, while there is no correlation found for interpretations, which is possibly due to the small amount of confusions for interpretations.

Furthermore we discuss the application of these measures to the evaluation of grammars. The measures can be used in two ways. In the design of a speech recognition grammar they can be used to determine the phonetic confusability of a grammar. For dialog system evaluation the measures can be used to estimate a grammar score based on samples from the grammar or recognized utterances, which can be added to other system parameters to evaluate a spoken dialog system [8].

	p	t	k
p	55.0	4.4	6.5
t	3.3	76.0	4.4
k	5.8	5.2	72.8

Table 2: *Percentages of phoneme confusion.*

2. Edit-distance based measures

The first distance we used is the standard minimum-edit-distance [1] or Levenshtein distance. We used an equal weight $\delta = 1$ for the edit operations substitution, insertion, and deletion to have a symmetric version of this distance metric. The minimum number of edit operations was normalized by the length of the words. This measure is not used in the evaluation, due to its poor performance on the number recognition data.

The second measure based on the minimum-edit-distance uses overlaps between articulatory phonetic features as a basis for adjusting the substitution costs of the edit distance as in [2].

From the overlap of phonetic features a phoneme similarity is derived. For computing the phoneme similarities a weighted Jaccard coefficient is used. It is defined as

$$\mathrm{jc}(p_1, p_2) = \alpha \frac{\mid X_1 \cap X_2 \mid}{\mid X_1 \cup X_2 \mid} \quad (1)$$

where X_1 and X_2 are the features of the phonemes p_1 and p_2 and α is a weight. Table 1 shows the phoneme features of the consonants *s* and *z*. The phoneme similarities computed according to Definition 1 are then used to adjust the substitution weights of the minimum-edit-distance. The more similar two phonemes are, the cheaper there substitution will be. The phoneme similarities range from 0 to 1.

The weighted phoneme similarities are used as substitution costs. The optimization of the weight α has a significant impact on the performance of the phonetic edit distance. The optimized α value for the number domain development test set is 0.6.

Among the most similar phonemes for this task were the consonants *z* and *s*. They appear for example in the beginning and end of *sechs (z E k s)* and have all but two features in common as shown in Table 1. According to Definition 1 the similarity between these two phonemes is therefore

$$\mathrm{jc}(z, s) = \alpha \frac{7}{11} = 0.6 \frac{7}{11} = 0.38\ . \quad (2)$$

A third distance measure is defined by using the standard minimum-edit-distance with perceptual similarities as weights for the substitution costs. The perceptual similarities are taken from [3] and are determined by a phoneme identification test. One advantage of this measure is that the self-similarity between phonemes needs not be 0, such that the distance between *sechs (z E k s)* and *sechs (z E k s)* is not necessarily 0. Table 2 shows some sample similarities for some phonemes. The optimized α value for the number domain development test set is 0.5.

3. HMM-distance based measures

As a fourth measure we used a distance measure that is based on the HMMs that are used by the recognizer. This measure is not used in the evaluation since we have no access to the acoustic models of the speech recognizer of the spoken dialog system. Here this measure is included to compare it with the optimized minimum-edit-distance measures.

The HMM measure is based on the Kullback-Leibler (KL) divergence between two Gaussian mixture models (GMM). To compute the KL divergence between two GMMs we use the method proposed in [9].

4. Optimization of distance measures and word confusion

The acoustic models used in these experiments were trained with HTK on the Speechdat-AT corpus [10] which contains audio data spoken by 1000 Austrian speakers both over mobile and landline telephones. Each speaker recorded 57 short utterances that are relevant to a command and control task. In order to obtain a useful number of confusions the experiments were carried out on the mobile phone data of the corpus which consist of 8kHz A-law audio. 39 dimensional PLP feature vectors including first and second order derivatives were extracted from the audio data. The acoustic models were state-tied Gaussian mixture HMMs with 12 mixtures per state and about 5000 states in total. The models were evaluated on continuous sequences of numbers ranging from 0 to 999 which were broken down into a vocabulary of 31 words. In order to focus entirely on the acoustic confusability, a context free grammar was used for the recognition experiments. This resulted in a WER of 18.27%.

Figure 1 shows the correlation between distances and word confusions for the four measures. Each bin shows the number of word confusions within this distance range, normalized by the total number of distances in this range. Thereby the distribution of the distance measures, which is not uniform is also taken into account.

Ideally there would be a high number of word confusions for low distances that is high similarities, and a low number of word confusions for high distances that is low similarities. Generally the distribution of three measures follows that rule, where the HMM distance performs best, followed by the articulatory and perceptual phonetic minimum-edit-distance, and the minimum-edit-distance.

The comparison of the minimum-edit-distance with the phonetic minimum-edit-distances shows that the latter outperform the former. In our experiments we saw that the optimization of the phonetic similarity weights (Definition 1) is essential. Without such an optimization the phonetic minimum-edit-distance performs worse than the measure without phonetic similarities.

5. Evaluation of measures and word confusion

In the evaluation we evaluated the two optimized phonetic minimum-edit-distance measures on data from a spoken dialog system.

The test data consisted of 1091 utterances from a spoken dialog system with a WER of 16.4%. The understanding error rate was however only 10.9%. This error rate was determined by comparing the natural language interpretation of the reference utterance with the one of the recognized utterance. If a gold standard for interpretations is applied, then the understanding error rate is 6.5%. In the first example in Table 3 the interpretation error disappears when using a gold standard that tells us that these two word strings must have the same interpretation. [] signifies that there is no interpretation.

To evaluate the performance of the measures for the test set we again estimated the correlation of word confusion and pho-

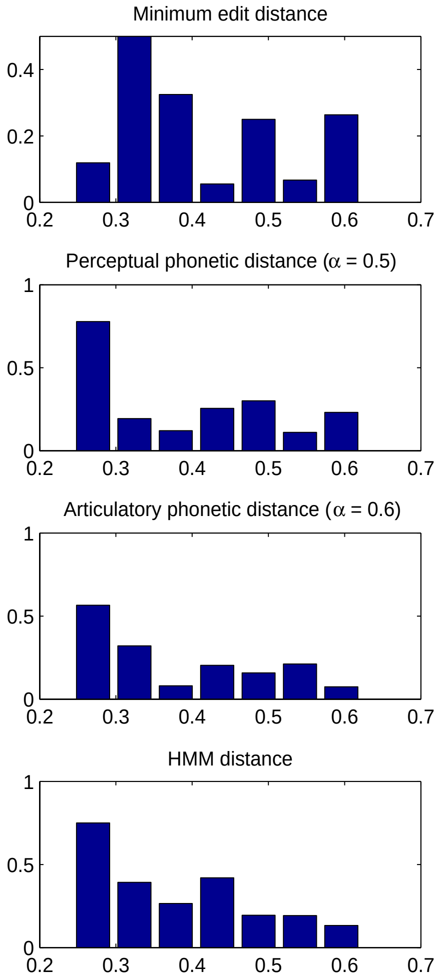

Figure 1: *Optimized correlation between distance measures and word confusion.*

netic distance of words. Additionally the correlation of natural language interpretation confusion and phonetic distance between natural language interpretations was estimated. The natural language interpretation is the concatenation of all slot values that are filled by an utterance. In Table 3 there are two examples for the slot “target”, which has the values “rechnung” and “aufträge”.

For measuring the phonetic distance between interpretations we take all utterances in the test data that lead to a certain interpretation (e.g. target-aufträge) and compute the phonetic distance to all utterances that lead to another interpretation. As the confusability between interpretations we take the mean or maximum values.

Figure 2 and 3 show the correlation between word confusions and articulatory and perceptual phonetic distances. It can be seen that most confusions are with low distances e.g. high similarities, which allows for the introduction of a threshold for confusability prediction.

	Words	NL interpretation
Rec. Ref.	NACHFRAGE zu aufträgen NACHFRAGEN zu aufträgen	target-aufträge []
Rec. Ref.	ABRECHNUNG ABBRECHEN	target-rechnung abbrechen

Table 3: *Word errors and understanding errors.*

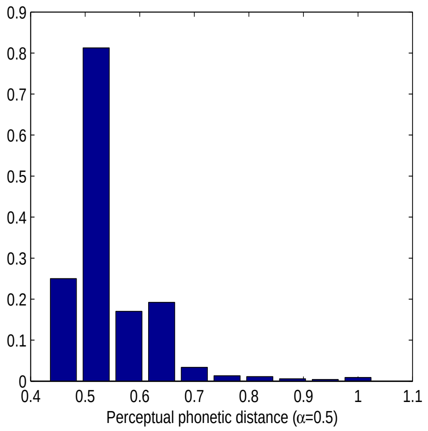

Figure 2: *Correlation between perceptual distance and word confusion.*

For the distances between interpretations and the confusion of interpretations no such correlation was found. This can be however due to the small amount of confusion data for interpretations. Another reason could be that the phonetic distance within interpretations can be high, depending on the phonetic distance of utterances that lead to the same interpretation.

6. Application to grammar optimization

We want to use the phonetic distance measures defined for words and interpretations for the analysis of spoken dialog system grammars. This shall help the grammar designer to optimize the grammars.

> S: *Wenn sie zurück zum Hauptmenü möchten sagen sie “abbrechen” (To return to the main menu say “quit”)*
> U: *Abbrechen (Quit)*
> S: *Haben sie Fragen zu ihrer Rechnung? (Do you have questions concerning charging?)*
> U: *Nein (No)*

In the above sample dialog taken from the test set *abbrechen/quit* is confused with the phonetically similar *Abrechnung/charging*, which leads to an annoying error. The same error is made multiple times. To find such possible sources of error the phonetic distance measures can be used.

For the analysis of spoken dialog systems we implemented four coherence measures that are based on the phonetic minimum-edit-distance measures. These are outer-grammar, inner-grammar, outer-vocabulary and inner-vocabulary coherence.

To compute the outer-grammar coherence we sample sentences from the grammar, put each sample into a class according to the interpretation it generates, and then we measure the

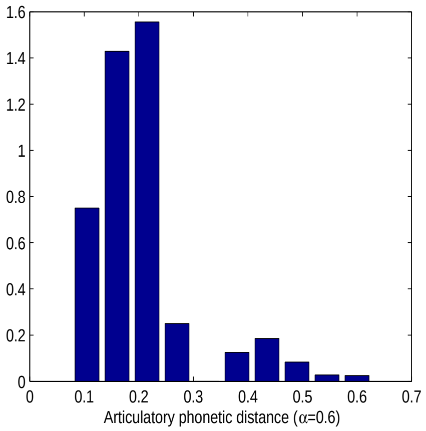

Figure 3: *Correlation between articulatory distance and word confusion.*

distance between interpretations as in Section 5. The distance between interpretations should be high to reduce the phonetic confusability.

The inner-grammar coherence measures the distance within an interpretation. The distance within an interpretation should be low. The vocabulary coherences are defined on the level of the vocabulary. Therefore sentences are sampled from the grammar. Then the vocabulary is extracted from the sentences, and added to a class according to the natural language interpretation of the sentence.

A graphical user interface allows the user to inspect lists of the most confusable interpretation pairs and the most confusable sentences/words between or within interpretations. One also can browse in a list of highly self-non-confusable interpretations, which means that the phonetic distance within the interpretation is high.

One gap between the evaluation of the methods on data from a spoken dialog system, and the use of these methods for the analysis of spoken dialog system grammars is the difference between language competence and language performance.

On the performance side (dialog system data) we see mostly short utterances. On the competence side (sampling sentences from grammars) we generally generate longer utterances. One important next step to apply these measures for the optimization of spoken dialog system grammars is therefore the pruning of the performance space.

7. Conclusion

We optimized two phonetic distance measures concerning their correlation to word confusion and compared them with a simple minimum-edit-distance measure and an HMM-based distance measure. The ordering of the measures showed that the HMM-based measure performs best followed by the minimum-edit-distance measures that use articulatory or perceptual phonetic information, followed by the simple minimum-edit-distance.

Then we applied the two optimized phonetic minimum-edit-distance measures to test data collected from logging data of a spoken dialog system. Here we could not use the HMM-based measure, since we had no access to the acoustic models of the speech recognizer. The correlation of the measures with word confusion and confusion of natural language interpretations was computed.

We saw that the optimized phonetic distance measures can be used for predicting the word confusion and thereby the WER by introducing a threshold. This is possible since most confusions happen with words that are phonetically similar to each other.

Finally we discussed the application of these measures for the analysis of spoken dialog systems. Therefore we defined different coherence measures that can be used for estimating within and between interpretation coherence on a sentence or word level.

8. Acknowledgements

This work was carried out within the project TIDE (Testbed for Interactive Dialog System Evaluation), which is a joint project of Deutsche Telekom Labs Berlin (T-Labs) and Telecommunications Research Center Vienna (ftw.). The authors also want to acknowledge the Siemens AG, Corporate Technology, Competence Center “User Interface Design” for providing the project team with the results of the usability test.

9. References

[1] Robert A. Wagner and Michael J. Fischer, “The string-to-string correction problem.,” *J. ACM*, vol. 21, no. 1, pp. 168–173, 1974.

[2] Stefan Schaden, “Evaluation of automatically generated transcriptions of non-native pronunciations using a phonetic distance measure,” in *Proceedings of LREC 2006*, Genova, Italy, 2006.

[3] A. Cutler, A. Weber, R. Smits, and N. Cooper, “Patterns of english phoneme confusions by native and non-native listeners,” *J. Acoust. Soc. Am.*, vol. 116, pp. 3668–3678, 2004.

[4] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi, “An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition,” *Bell Syst.Tech.J.*, vol. 62, pp. 1035–1074, 1983.

[5] B.H. Juang and L.R. Rabiner, “A probabilistic distance measure for hidden markov models,” *ATT Technical Journal*, vol. 64, February 1985.

[6] M. Falkhausen, H. Reininger, and D. Wolf, “Calculation of distance measures between hidden markov models,” in *Proc. Eurospeech*, 1995.

[7] Claus Bahlmann and Hans Burkhardt, “Measuring HMM similarity with the Bayes probability of error and its application to online handwriting recognition,” in *Proc. of the 6th ICDAR*, 2001, pp. 406–411.

[8] Sebastian Möller, *Quality of Telephone-Based Spoken Dialogue Systems*, Springer, New York, 2004.

[9] Jacob Goldberger and Hagai Aronowitz, “A distance measure between GMMs based on the unscented transform and its application to speaker recognition,” in *Proceedings of INTERSPEECH-2005*, 2005, pp. 1985–1988.

[10] M. Baum, G. Erbach, and G. Kubin, “Speechdat-AT: A telephone speech database for Austrian German,” in *Proceedings of the LREC workshop on Very Large Telephone Databases (XL-DB)*, Athens, Greece, 2000.