AUTOMATIC PHONETICS-DRIVEN RECONSTRUCTION OF MEDICAL DICTATIONS ON MULTIPLE LEVELS OF SEGMENTATION

Stefan Petrik and Franz Pernkopf

Signal Processing & Speech Communication Laboratory,
Graz University of Technology, Graz, Austria
stefan.petrik@tugraz.at, pernkopf@tugraz.at

ABSTRACT

Automatic phonetic reconstruction of medical dictations from non-literal and automatically recognized speech transcripts leads to closer-to-literal transcripts for training. In this paper, we introduce an extended alignment method assessing multiple levels of text segmentation and show how open issues like wrong segmentation in the recognized transcript can be resolved. Furthermore, the effect of context-dependent reconstruction and the phonetic similarity threshold on the quality of the reconstructed transcription is measured. Experiments show an increase in precision between 0.7% and 4.7% absolute without loss in recall for the combined system incorporating all of these techniques in comparison to the system in [1].

Index Terms: Automatic transcription, phonetic similarity, text alignment, syllabification, dictation

1. INTRODUCTION

Literal transcriptions of spoken input are a valuable resource for training the acoustic and language models in automatic speech recognition (ASR). For large vocabulary continuous speech recognition (LVCSR) systems, human transcriptions are expensive and time consuming. Automatic generation of literal transcriptions is thus a desirable, but challenging task.

In medical dictation systems, large amounts of non-literal transcriptions of spoken input, produced by trained typists are available in the form of medical reports. In contrast to literal transcriptions, these do not accurately represent spoken input because of inherent differences between spoken and written language like filled pauses, self-corrections, short forms, etc. Furthermore, medical reports are produced to conform to a standardized, written form. The original utterance has possibly been reformulated or restructured by the typist as shown in the following example:

she basically lays in bed non-responsive	*(spoken)*
she basically lays in bed not responsive	*(recognized)*
Basically she is nonresponsive.	*(written)*

Previous work on using non-literal transcripts for training can be found for the domain of academic lectures [2], closed captions in broadcast news [3], predicting ASR errors [4], and also medical dictations [5]. In a previous paper [1], we presented a phonetic similarity measure for matching automatically recognized transcripts and non-literal medical reports to reconstruct a literal transcription of a medical dictation. Mismatches between the texts were either classified as corrected ASR errors, assuming that ASR errors are phonetically similar, or possible reformulations inserted by the typist in case of phonetic dissimilarity. The simple reconstruction based on this classification showed significant improvements on the word error rate in comparison with a reference transcription, but was not able to deal with certain types of errors, like wrong segmentations in the recognized texts, or massive reductions due to fast speech.

This paper presents two main improvements to the initial approach. First, a refined text alignment that explores multiple levels of text segmentation to better handle alignment mismatches caused by segmentation errors in ASR. Second, a comparison of a rule-based and a data-driven text reconstruction step that interprets the enhanced alignment, before it generates a reconstruction hypothesis. In an experimental study, we demonstrate the effects of the enhanced alignment with rules which use context and overlap information on syllable level and automatic classifiers operating with the same features. Furthermore, the influence of the phonetic similarity threshold is shown. Based on the experimental results, we discuss the contribution of each implemented technique and conclude the paper with an outlook for future work. In the following, we will refer to the automatically recognized draft transcription as the *recognized text*, to the manually corrected medical report as the *written text*, and to the literal reference transcription of the dictation as the *reference text*.

2. LEVELS OF TEXT SEGMENTATION

In general, text alignment is done for a complete recognized respectively written text document - at the *document level*. The following levels of text segmentation were selected primarily based on observations in data and not on the basis of linguistic knowledge:

- **Mismatch regions:** We define a *mismatch region* (ERR) as a sequence of the edit operations insertion (INS), deletion (DEL), and substitution (SUB) in a word level alignment. This definition is helpful, since actual mismatches can be composed of several adjacent mismatch edit operations as shown in table 1. Assuming that the word order in the aligned documents is not deviating too much, focusing on the mismatch regions is an appropriate simplification of the alignment task. To ensure that a mismatch region was not split, we allow a mismatch region to be interrupted by at most one identity edit operation (COR).
- **Words / short phrases**: Text alignment is expressed by edit operations at *word level*. The segmentation is defined by the ASR lexicon and final automatic formatting applied to the recognized text. Apart from lexical words, short phrases are included as well

This research has been carried out in the context of the projects SPARC and COAST. We gratefully acknowledge funding by the Austrian FIT-IT and KNet programs, ZIT Zentrum fuer Innovation und Technologie, Vienna, Steirische Wirtschaftsfoerderungsgesellschaft mbH and Land Steiermark. http://www.sparc.or.at, http://www.coast.at.

in terms of formatted entities like dates, times, physical values (= number + physical unit), or determiner phrases like 'the patient'.

- **Syllables**: We observed in our data that segmentation errors in recognized texts were more likely to occur at syllable boundaries than at any other position within a word. This observation is in accordance with previous studies [6] which found the *syllable level* to be more robust against variation in conversational speech than the phoneme level.
- **Phonemes**: Phonetic similarity matching itself is done at *phoneme level*, the lowest matching level. The sequence of phonemes which produced the recognized text is compared to an automatic phonetic transcription of the written text.

3. TEXT ALIGNMENT AND RECONSTRUCTION

The reconstruction task is a three-step process. First, recognized and written text are aligned with standard Levenshtein alignment [7] to detect mismatch regions. Then, the mismatch regions are re-aligned, to ensure that correspondences between texts are correctly labelled. Finally, a reconstruction hypothesis is generated by applying reconstruction rules to the previously calculated text alignment.

3.1. Phonetic similarity matching

Both, re-alignment and reconstruction are based on phonetic similarity matching [1]. A phonetic scoring function, defined on a stochastic string edit distance model is used to compute a similarity score of two phoneme strings. In the stochastic model, similarity is defined as the negative log-likelihood of the joint probabilities for the input strings. The scoring function is normalized with respect to the input string length and a score scale of 0.0 (dissimilar) to 10.0 (identical). The similarity decision is made with an adjustable threshold t.

3.2. Automatic syllabification

An annotated expert phonetic lexicon for the highly specific vocabulary used in the medical domain was not available for our data. Therefore, an automatic syllabification algorithm was used to determine syllables from words online [8]. The algorithm is based on Optimality Theory (OT), where phonological processes are modelled by applying ranked constraints on base forms to obtain surface forms. As primary stress information was not available in the phonetic lexicon, the algorithm had to be modified. The modification degraded the performance of the algorithm in terms of accuracy of the syllable boundaries, but not the number of detected syllables. In an informal test, it still returned correct results for 80 out of 100 words.

With this algorithm, the word level units for recognized and written text are split into sequences of syllables. The alignment algorithm is then applied recursively on the syllable sequences. Adjacent words are not only aligned, but also tested for overlap on syllable level. The word level alignment label is therefore replaced by an overlap symbol string. The resulting alignment expresses both, word and syllable level correspondences. Consider the sample alignment in table 1. Within the first mismatch region, the word `Charcot` was incorrectly recognized and split into `sharp` and `cold`. The syllable level alignment, however, shows that `sharp` corresponds to the first, and `cold` to the second syllable of `Charcot`. As syllable alignment is determined based on phonetic similarity, the alignment may sometimes look confusing. The short words `of` and `in` are not aligned with each other, since `in` is phonetically more similar to the last syllable of `ulceration` than to `of`.

3.3. Rule-based text reconstruction

The hypothesized literal transcription is generated with the help of reconstruction rules. These rules define which parts of the recognized or written text are selected to appear in the reconstructed text. Each rule is defined for a sequence of alignment labels. Whenever the label sequence appears in the alignment, the rule applies and an action specified by the rule is executed. For better control, the rule action can be conditioned on e.g. phonetic similarity between the matched words. Since more than one rule can match for a certain sequence of alignment labels, rules match on a first-come first-serve basis, meaning that rule precedence has an effect on the result. In the experiments, results were only given for rule orderings which gave the best performance.

To test the effects of the previously described techniques, we defined the following reconstruction rules, where an alignment label is either the identity edit operation (COR) or a sequence of syllable overlap symbols [=, <, >] (c.f. table 1):

- **Baseline**: only identical words in the alignment (COR) are reconstructed, mismatch regions are ignored.
- **Recognized-only (REC)**: for each alignment label, select the *recognized text* for reconstruction.
- **Written-only (WRI)**: for each alignment label, select the *written text* for reconstruction.
- **Phonetic similarity (PHO)**: for alignment labels containing matching syllables (=), select *written text*, if phonetic similarity is higher than a threshold value, and *recognized text* otherwise. REC and WRI are the extreme cases, if the similarity threshold value is maximum or zero.
- **Context (CTX)**: for sequences of 1, 2, or 3 alignment labels containing at least one syllable match (=), select *written text*, if phonetic similarity is higher than a threshold value. The idea behind this rule is that longer corresponding regions in the alignment are more likely to be real correspondences.
- **Overlap, greedy (OVG)**: for sequences of 2 or 3 alignment labels, where inserted or deleted syllables (</>) are either preceded or succeeded by at least one matching syllable (=), select *written text*, if phonetic similarity is higher than a threshold value. This rule collects all word sequences showing any possible overlap at syllable level without regard of the matching order.
- **Overlap, selective (OVS)**: for sequences of 2, 3, or 4 alignment labels, where matching syllables (=) are first succeeded by inserted (<), and then preceded by deleted (>) syllables, select *written text* if phonetic similarity is higher than a threshold value. This pattern is typical for segmentation errors in the recognized text.

Table 1 illustrates the effect of each rule on a sample alignment. The **phonetic similarity** rule resolves each alignment line on its own and therefore fails in both mismatch regions. The **context** rule, however, performs much better, as it is activated whenever a group of matching syllables appears. Still, it is not enough as it does not handle stand-alone insertions or deletions appropriately. The **greedy overlap** rule can handle insertions and deletions whenever they appear in terms of a syllable overlap. However, it is not activated when there is a direct match (`though` ↔ `no`). The **selective overlap** rule, finally, matches only the precise first segmentation error, where the syllable counts exactly match. Accidental matches are therefore impossible. This example indicates that combination of rules may be beneficial.

3.4. Data-driven text reconstruction

For data-driven text reconstruction, we use different classifiers to produce the hypothesized literal transcription which is the 2-class

written text	↔	*recognized text*	**PHO**	**CTX**	**OVG**	**OVS**	*reference text*
a	COR	a	a	a	a	a	a
Char·cot	=<	sharp	Charcot	Charcot	Charcot	Charcot	Charcot
	>=	cold	cold				
foot	=	foot	foot	foot	foot	-	foot
,	<		-	-	,	-	
though	=	no	though	though	-	-	though
there is	COR	there is	there is	there is	there is	there is	there is
no	COR	no	no	no	no	no	no
ul·ce·ra·tion	====	al·te·ra·tion	ulceration	ulceration	ulceration	-	ulceration
	<<<=	in	in			-	
of	<		-	-	of	-	of
skin	COR	skin	skin	skin	skin	skin	skin

Table 1. A sample alignment containing two mismatch regions (dashed boxes), together with reconstruction rule results. Syllable boundaries are marked with dots [·]. Note that the [=]-overlap symbol just indicates correspondence, not equality of syllables, in contrast to the insertion [<] and deletion [>] symbols which label non-matching syllables. The solid boxes highlight lines affected by each rule, dashes [−] mark parts not covered by the rule.

output of a classifier, i.e., either *written text* or *recognized text*. For classifier training, the class labels are produced by aligning the reference text with the written text. The features are derived from the automatic alignment and the phonetic similarity score, computed for the aligned written and recognized phoneme strings. In addition, this score is derived for 3 consecutive phoneme strings to model the dependency of adjacent words in the classifier. The remaining features are computed from the sequence of syllable symbols ([=], [<], [>]). Therefore, the sequence is split into 3 equal parts. After assigning values to the symbols ([=] : 0,[<] : -1,[>] : 1), the mean and standard deviation of each part serve as feature. The last feature employed denotes the length of the syllable symbol sequence. In sum, 9 features are used with the following classification approaches [9]:

- ***k*-NN**: k-nearest neighbor classifier. For table 2, $k = 9$.
- **NN**: Neural network with 3 layers. The number of neurons in the input and output layer is set to the number of features and the number of classes, respectively. The number of neurons in the hidden layer is set to 70. We use Levenberg-Marquardt backpropagation for training, a hyperbolic tangent sigmoid transfer function for the neurons in the input and hidden layer, and a linear transfer function in the output layer.
- **SVM**: The support vector machine with the radial basis function (RBF) kernel uses two parameters C^* and σ, where C^* is the penalty parameter for the errors of the non-separable case and σ is the parameter for the RBF kernel. We set the values for these parameters to $C^* = 1$ and $\sigma = 1.5$.

The optimal choice of the parameters, kernel function, number of neighbors, and transfer functions of the above mentioned classifiers has been established during extensive experiments. Five-fold cross-validation is used to produce the results with the classifiers. Throughout our experiments, we use exactly the same data partitioning for each training procedure.

4. EXPERIMENTS

The reconstruction rules were tested on an evaluation corpus by using them for the reconstruction of a literal transcription. The evaluation corpus consisted of 735 written and recognized texts of about 335.000 words, as well as manual reference texts for validation of the hypothesized reconstruction. The texts were selected such that they equally represent three ranges of average word error rates (WER) for the recognized text. Hesitations and incomplete words were removed beforehand to avoid biased results.

	5-13% WER			20-25% WER			40-45% WER		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Baseline	100	78.9	88.2	99.9	64.6	78.4	99.5	46.0	62.9
REC	83.3	**93.4**	88.1	79.0	85.7	82.2	66.7	71.9	69.2
WRI	92.8	93.1	92.9	89.9	**89.6**	**89.8**	85.9	**85.4**	**85.6**
PHO	**96.3**	91.6	**93.9**	**93.1**	84.5	88.6	**87.4**	72.5	79.3
CTX	97.6	90.4	93.8	95.4	82.8	88.7	93.1	69.8	79.8
OVG	97.9	86.4	91.8	95.8	78.3	86.2	93.1	65.7	77.0
OVS	**99.8**	79.5	88.5	**99.6**	65.6	79.1	**98.8**	47.3	64.0
CTX+OVG+OVS	97.0	**91.1**	**94.0**	94.7	**84.3**	**89.2**	92.1	**72.6**	**81.2**
k-NN	**94.9**	92.8	93.8	**91.6**	87.9	89.7	**87.0**	83.1	85.0
NN	**94.9**	**93.0**	**94.0**	91.2	88.4	89.7	86.5	83.7	85.1
SVM	94.8	92.9	93.9	91.2	**88.5**	**89.8**	86.4	**84.1**	**85.3**

Table 2. Reconstruction results in % for rule-based approach (first block) and data-driven approach (second block). Best results for each row grouping are boldface.

In earlier experiments, results on this task were reported in terms of word error rates between the hypothesized reconstruction and a manual reference text. Word error rate, however, combines wrongly hypothesized and missed words in one measure, making it difficult to optimize reconstruction rules. The retrieval power is better expressed in terms of the information retrieval metrics $Recall = \frac{|COR|}{|COR|+|MISS|}$, $Precision = \frac{|COR|}{|COR|+|WRONG|}$, and their harmonic mean $F1$, where |COR| is the number of reconstructed words with perfect correspondence in the reference text, |MISS| is the number of words in the reference text without correspondence in the reconstructed text, and |WRONG| is the number of reconstructed words without correspondence in the reference text. The obtained *Recall/Precision/F1* scores for the systems described in 3.3 and 3.4 are shown in table 2.

The first block of table 2 are rule-based approaches. The first three systems (REC, WRI, PHO) are the same systems as used in [1], so results are comparable. The CTX, OVG, and OVS systems explore context and syllable overlap, and CTX+OVG+OVS is the combination of these rules. The last block lists the data-driven systems k-NN, NN, and SVM in comparison to the rule-based systems.

Finally, we tested the impact of the similarity threshold value on the final system incorporating all rules. The threshold value can be adjusted between $t = 0.0$ (no similarity) and 10.0 (identity) and was varied from $t = 5.0$ to 10.0 in the experiments. The resulting curves are plotted in a Recall/Precision diagram, shown in figure 1.

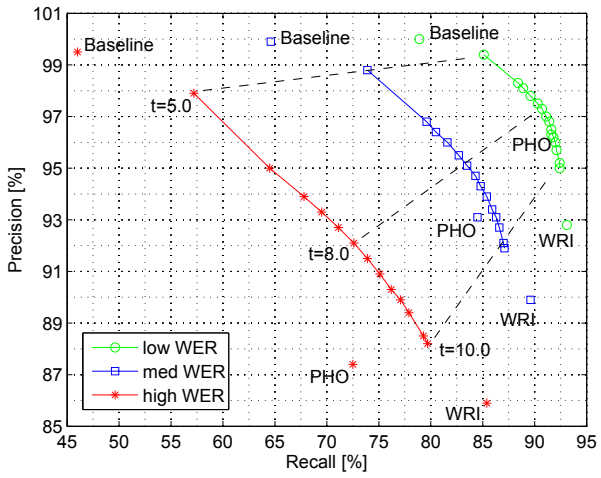

Fig. 1. Recall/Precision diagram derived from the CTX+OVG+OVS system by varying the phonetic similarity threshold t between $t = 5.0$ and $t = 10.0$ for high, medium, and low WER texts (c.f. table 2).

5. DISCUSSION & OUTLOOK

The recognized text (REC) is not a good starting point for reconstructing a literal transcription. Although the recall scores are comparable with the other methods, many errors are taken over from the recognition process, resulting in poor precision. The written text (WRI) is more reliable for the domain of medical dictations. Phonetic similarity matching (PHO) further boosts the precision, as it only selects those text parts with corresponding acoustic evidence.

Using context (CTX) in the phonetically controlled reconstruction improves the precision at the expense of the recall, as the F1 scores remain almost the same. Only for the high WER case, a small gain of 0.5% absolute in the F1 scores can be observed. However, this means that further rules have to be added to raise the recall.

The greedy exploration of overlap on syllable level (OVG) returned surprisingly precise results which are absolutely comparable to using contextual information. This applies even more to the selective overlap rule (OVS), which has only very little gain in recall in comparison to the baseline, but almost maximum precision. These findings indicate that the combination of these rules could be beneficial. The combination of all rules shows the best performance for all WER ranges. In comparison to the simple approach (PHO), there is a gain in precision without loss of recall.

Finally, optimizing the threshold value for phonetic similarity also contributes to the overall performance. The trade-off between recall and precision is not linear, as the graphs in figure 1 show. The best recall/precision value pairs were obtained for a similarity threshold value $t = 8.0$, independent from the initial WER.

The data-driven systems are closer to the written text only (WRI) reconstruction than the rule-based system, showing improvement in precision for all WER ranges. The rather simple *k*-NN classifier consistently produces the highest precision while the more complex NN and SVM classifiers achieve higher recall scores. The rule-based system outperforms the data-driven system only for low error rates. Still, the main benefit of the data-driven approach is that no handcrafting of rules and no phonetic similarity threshold is required.

The impact of this text reconstruction method is currently being tested for re-training of an ASR language model. For this application, one would intuitively focus on high precision of the reconstructed texts to accurately model the specifities of spoken language. The particular trade-off between precision and recall still has to be studied to give recommendations for selecting the optimum operating point for the reconstruction system. Furthermore, transferring the method to other domains, e.g. enhancing imprecise closed captions will also be an interesting future application.

6. CONCLUSION

We presented a sophisticated automatic transcription system for medical dictations based on phonetic similarity matching and text alignment on multiple levels of segmentation. For comparison, these features are implemented in a rule-based and an automatic classification reconstruction system. We investigated the effect of syllabification in text alignment, context in text reconstruction, and the phonetic similarity threshold on the performance of the system.

The usage of multiple reconstruction rules with high precision, but lower recall like context- or syllable-based rules returned more accurate results than the data-driven approach, except for high initial word error rates. The combination of all the described techniques improved the precision of the reconstructed literal transcription between 0.7% and 4.7% absolute in comparison to a simpler system, also based on phonetic similarity matching.

7. ACKNOWLEDGEMENTS

We would like to thank our partners Johannes Matiasek, Alexandra Klein, Jeremy M. Jancsary, Martin Huber, and Harald Trost at the Austrian Research Institute for Artificial Intelligence (OFAI) and Christina Drexel at Philips Speech Recognition Systems for providing software and for many valuable discussions.

8. REFERENCES

[1] Stefan Petrik and Gernot Kubin, "Reconstructing medical dictations from automatically recognized and non-literal transcripts with phonetic similarity matching," in *Proc. ICASSP*, Honolulu, Hawaii, 2007, pp. 1125–1128.

[2] Timothy J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Proc. ICSLP*, Pittsburgh, Pennsylvania, 2006, pp. 1606–1609.

[3] Lori Lamel, Jean-Luc Gauvain, and Gilles Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, pp. 115–129, 2002.

[4] Eric Fosler-Lussier, Ingunn Amdal, and Hong-Kwang Jeff Kuo, "A framework for predicting speech recognition errors," *Speech Communication*, vol. 46, pp. 153–170, 2005.

[5] Sergey Pakhomov, Michael Schonwetter, and Joan Bachenko, "Generating training data for medical dictations," in *Proc. NAACL*, Pittsburgh, Pennsylvania, 2001.

[6] Steven Greenberg, "Speaking in shorthand: a syllable-centric perspective for understanding pronunciation variation," *Proc. ESCA Workshop 'Modeling Pronunciation Variation For Automatic Speech Recognition'*, pp. 47–56, 1998.

[7] Vladimir Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals.," *Soviet Physics - Doklady*, vol. 10, pp. 707–710, 1966.

[8] Mike Hammond, "Syllable parsing in English and French," Rutgers Optimality Archive, 1995, http://roa.rutgers.edu/.

[9] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.