RECONSTRUCTING MEDICAL DICTATIONS FROM AUTOMATICALLY RECOGNIZED AND NON-LITERAL TRANSCRIPTS WITH PHONETIC SIMILARITY MATCHING

Stefan Petrik and Gernot Kubin

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria
stefan.petrik@tugraz.at g.kubin@ieee.org

ABSTRACT

In this paper, we describe the automatic reconstruction of literal transcriptions for medical dictations from a non-literal transcription and an automatically recognized speech transcript by phonetic similarity matching and alignment. We present a customized phonetic similarity measure which is trained on a set of phonetically similar string pairs, returns interpretable alignment results, and is robust in its application. Furthermore, we introduce flexible automatic phonetic transcription with regular expressions to deal with formatted entities in written texts and alternative pronunciations in recognized texts. In an evaluation, our method reduced the word error rate for the reconstructed transcription by 12% relative.

Index Terms— String edit distance, trained similarity measure, phonetic similarity, Levenshtein distance, dictation

1. INTRODUCTION

In automatic speech recognition (ASR), literal transcriptions of spoken input are needed for training the acoustic and language models of the recognizer. Such transcriptions are, however, costly in production, as considerable efforts by trained, human transcribers are required. The amount of such literal transcriptions needed for speaker-independent, large vocabulary continuous speech recognition (LVCSR) makes this an expensive and time-consuming task.

However, in medical dictation systems, 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, etc. Furthermore, medical reports are produced to conform to a standardized, written form meaning that the original utterance has possibly been reformulated or restructured by the typist as shown in the following example:

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

To make use of the large amount of data collected every day, we present a method for automatically reconstructing a transcription which is closer to a literal one than an automatically recognized or non-literal one. Our approach is based on phonetic similarity matching applied to large corpora of paired automatically produced draft transcriptions and manually edited medical reports. We classify mismatches between these texts as either corrected ASR errors, assuming that ASR errors are phonetically similar, or possible reformulations inserted by the typist in case of phonetic dissimilarity. According to this classification, the corrpesponding word from either the automatically recognized or the medical report is selected and an enhanced transcription can be composed.

For medical dicatations, this reconstruction task was already described in [1]. There, the authors proposed an augmented probabilistic finite state model for generating a semi-literal transcription. In [2], transcription generation was presented for recorded academic lectures with a finite state transducer approach. Phonetic similarity matching has been used for tasks like modeling pronunciation variation [3], predicting ASR errors [4], or information retrieval [5].

In the following, we will refer to the automatically recognized draft transcription as the *recognized text*, and to the manually corrected medical report as *written text*. First, we describe the available text corpora of recognized and written texts. Then we present a customized phonetic string edit distance measure, which combines an advanced automatic phonetic transcription tool, featuring pronunciation variant generation and regular expression syntax with a trainable string edit distance measure. In an evaluation, we demonstrate the benefit of this measure and conclude the paper with a discussion of the results and an outlook for further research.

2. DATA DESCRIPTION

For reconstruction, only the mismatching parts of an alignment between the two texts are of interest, as this task is trivial for matching parts. Generally, mismatches between texts on word-level are described in terms of the mismatch edit operations insertion (INS), deletion (DEL), and substitution (SUB). This way, a word error rate can be determined easily, but mismatch interpretation is difficult since actual mismatches can be composed of several adjacent mismatch edit operations as shown in the example below. For this reason, a *mismatch region* (ERR) is defined as a contiguous sequence of mismatching edit operations such as to preserve correspondences between matched words.

(written)		(recognized)
left-to-right	SUB	left
	INS	to
	INS	right
	↓	
left-to-right	ERR	left to right

This research was carried out in the context of the SPARC project, a joint project by the Signal Processing and Speech Communication Laboratory at Graz University of Technology, the Austrian Research Institute for Artificial Intelligence (OFAI), and Philips Speech Recognition Systems. SPARC is funded by the FIT-IT program of the Austrian Federal Ministry for Transport, Innovation, and Technology under contract nr. FIT-IT-809 258. For further information, see http://www.sparc.or.at

A statistical study of a corpus of 80.000 medical reports with 38 million words revealed an average length of 2.3 words for a mismatch region and an average occurrence of 3.6 times for this region within the corpus. Regions occurring only once account already for 60% of all mismatches while frequent regions occurring $\geq$ 10.000 times only account for about 11% of all mismatches. Such highly frequent mismatches are e.g. insertions or deletions of punctuations and short words. On the other hand, regions of length 1 cover around 20% of all mismatches, and 75% of all mismatches occur in regions of length ≤ 5. For the reconstruction task, this means that only relatively short symbol sequences have to be matched.

Mismatches are introduced by the dictating person, automatic speech recognition, and the human transcription process itself. The dictating person speaks freely in general, thus hesitations, self-corrections, and repetitions can be observed quite often in the recordings, but of course not in the medical reports. Automatic speech recognition itself is error-prone, resulting in the confusion of words which are phonetically similar. The transcription process completes the range of mismatch sources by adding formatting to the text according to previously defined standards. Formatting affects the text in two ways: First, by additional structure like inserted punctuations, paragraph breaks, or capitalization of words, and second, by formatting of particular document entities like numbers, dates, times, quantities, etc. The latter formatting step makes reconstruction difficult, as different speaking variants are mapped onto a standardized written form. Furthermore, the structure and style of the text can be altered by reformulations of the typist as well. These alterations include expansion of abbreviations, acronyms, and short forms, or grammatical corrections like changes in genus, tempus, or numerus to put the final written text into a proper stylistic and grammatical form.

Except for the last mentioned reformulations, these mismatches can be tackled by phonetic analysis and similarity matching. In the following, a phonetic similarity measure is presented which is customized to dictated texts.

3. A CUSTOMIZED PHONETIC SIMILARITY MEASURE

Before a literal transcription can be reconstructed, it is essential to have an accurate alignment of the mismatch regions. This is achieved by an alignment procedure which is based on similarity measurement. The following phonetic scoring function is part of the alignment framework presented in [6], where the scoring function is separated from the actual dynamic programming alignment algorithm. A schematic view of the scoring function is depicted in figure 1. The scoring function uses three text resources for comparison: the phonetic symbol sequence from the recognized text, the orthographic word sequence from the recognized text and the word sequence from the written text. The two main components for similarity matching are explained in more detail below.

3.1. Automatic phonetic transcription (APT)

In a first step, the written text is transferred to the phonetic domain with automatic phonetic transcription (APT). This is done by a simple lexicon lookup. The used phonetic lexicon contained 160.000 words with 197.000 pronunciations. It included common as well as domain-dependent vocabulary and was compiled from customary and publicly available ressources like CMUdict [7]. To improve coverage on formatted text parts, a de-formatting grammar is applied to formatted text units. The de-formatting grammar is an inverted version of a formatting grammar used in the speech recognizer which now produces speaking variants for a given formatted entity as shown in the following example:

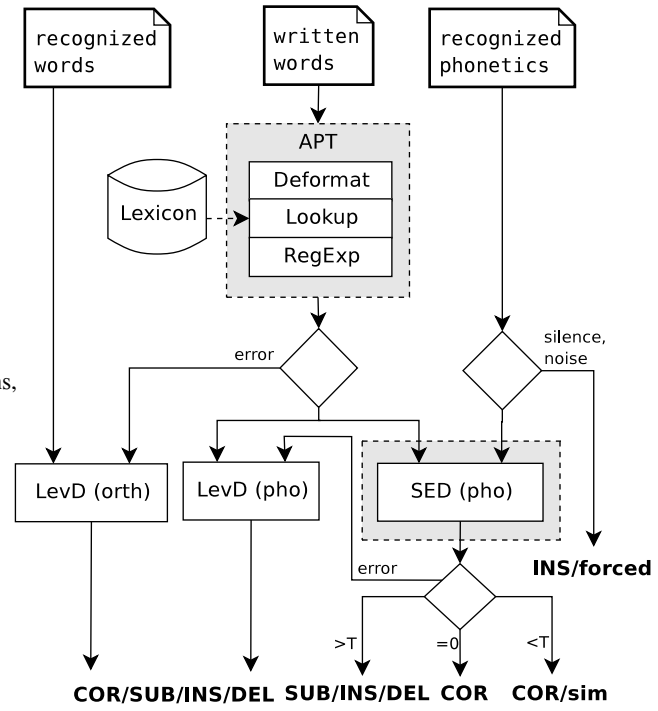

Fig. 1. Block scheme of phonetic similarity function: automatic phonetic transcription (APT), trainable string edit distance measure (SED), and Levenshtein measure (LevD)

```
December 6  →  December the sixth
               December O six
               sixth of December ...
```

Furthermore, a simple regular expression syntax was defined to encode the possibly many speaking and pronunciation variants in a single string. The extended syntax allows grouping and OR-ing of expressions as described in the corresponding BNF grammar:

$$\begin{aligned} expr &:= group^{+} \\ group &:= \text{"("}\ word^{+}\ (\text{"|"}\ word^{*})^{*}\ \text{")"} \\ word &:= [A..Z, a..z] \end{aligned}$$

Since the succeeding word after the | operator is optional, whole words can also be omitted. This is particularly useful for dealing with hesitations or dictated formatting instructions which do not appear in the written text by definition.

The recognized text still contains non-speech events like silence or noise markers which do not have a phonetic transcription and which are by definition not contained in the written text. These parts get assigned a score which automatically forces them to be marked as an insertion. After that, it is certain that the remaining string pairs are valid phonetic strings that can be fed to the phonetic similarity matching model. Whenever the APT fails, it is not possible to do phonetic matching, so the string pair can only be matched in the orthographic domain with the Levenshtein measure (LevD) [8].

3.2. Trainable string edit distance measure (SED)

The main component of the phonetic scoring function is a trainable string edit distance measure based on the stochastic model presented in [9]. In this model, a string pair $\langle x, y \rangle$ is represented by all sequences of edit operations z_i which produce that pair. Assuming that each pair can be produced by at least one edit sequence, the probability of the pair is then the sum of the probabilities of all edit sequences for that pair.

$$p(x, y|\theta) = \sum_{\{z^n\#:v(z^n\#)=\langle x,y\rangle\}} p(z^n\#|\theta) \ , \qquad (1)$$

where # is the sequence termination symbol and $v(z^n\#)$ defines the set of all terminated edit sequences producing $\langle x, y \rangle$. Since every z_i has a probability $p(z_i)$ assigned and the model is memoryless, $p(z^n\#|\theta)$ is the product of the probabilities of the single edit operations. These probabilities $p(z_i)$ are learned from a corpus of predefined, similar string pairs with an EM algorithm [9]. Accumulating the probabilities for all edit sequences, a similarity measure can now be defined as

$$d(x, y) = -\log \ p(x, y|\theta) \ . \qquad (2)$$

Two issues should be noted at this point. First, the similarity value decreases exponentially with the input string length due to the usage of the distinct termination symbol #. Therefore, the similarity value needs to be normalized by the sum of the input string lengths. Furthermore, the similarity measure is never zero since each edit operation has assigned a probability $0 < z_i < 1$. To still be able to detect exact matches, the systematic bias is subtracted symmetrically to normalize the measure to zero according to the following formula:

$$d_0(x, y) = d(x, y) - \frac{1}{2} \cdot [d(x, x) + d(y, y)] \qquad (3)$$

Prior to matching, the regular expressions generated by the automatic phonetic transcription have to be expanded again, as only the minimum score for all possible realizations is returned. Phonetically highly similar pairs with a score below a heuristically set threshold *T* are labelled with a separate *COR/sim* tag to distinguish them from other substitutions in the alignment. Finally, in case the stochastic model fails, another fallback to the Levenshtein measure is done, this time with phonetic strings.

The model was trained in three EM iterations with a small set of 13383 string pairs obtained from manual narrow phonetic transcriptions of the evaluation corpus. For each word in the transcription, a string pair consisting of the canonical transcription obtained from the phonetic lexicon and the actual phonetic transcription was compiled. This way, phonetic similarity is clearly defined and frequent phoneme confusions can be learned easily from real-world data.

Figure 2 displays the learned parameter distribution. As expected, most of the probability mass was assigned to identity operations (main diagonal). Furthermore, vowels were likely to be substituted by schwa and vice versa. Voiced-unvoiced substitutions between `t` and `d` were also quite prominent, just like substitutions between the syllabic (`n=`, `m=`, `l=`) and non-syllabic forms (`n`, `m`, `l`) of the semi-vowels. The learned parameter set clearly reflects the phonetic knowledge that can be observed in dictated speech.

4. EXPERIMENTS

The phonetic scoring function was tested on an evaluation corpus by using it for the reconstruction of a literal transcription. The reconstruction was done according to the following simple decision rules

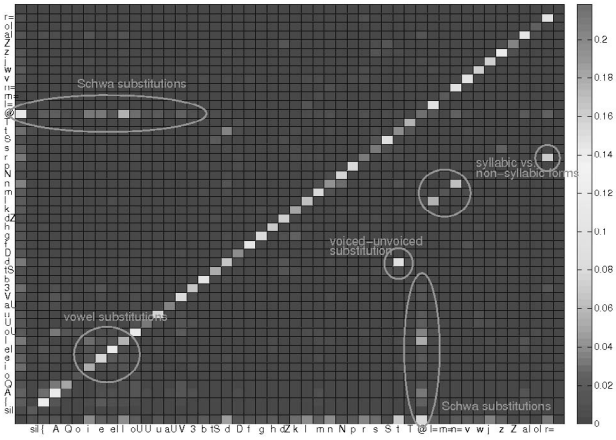

Fig. 2. Probability distribution for parameters z_i after 3 EM iterations. Phonetic symbols are in SAMPA notation.

applied to the calculated alignment tags:

COR	→	recognized text
COR/sim	→	written text
SUB, INS	→	recognized text
DEL, INS/forced	→	-

For full matches (COR), the recognized text (which then is identical to the written text) was chosen for reconstruction. Phonetically highly similar words (COR/sim) were hypothesized as corrected recognition errors, so the written text was used for reconstruction. Phonetically dissimilar words (SUB) or inserted words (INS) like hesitations were taken from the recognized text, assuming that the written text was maybe reformulated at this position. Deleted words (DEL) and forced insertions (INS/forced) were not included in the reconstruction. Finally, the hypothesized reconstructed text (HYP) was compared to the reference literal transcription and the word error rate computed with standard Levenshtein alignment.

The evaluation corpus consisted of 735 written and recognized texts of about 335.000 words, as well as reference manual transcriptions 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. Furthermore, the corpus was also divided based on the number of words included in the reference literal transcription. The obtained average word error rates for reconstruction are shown in table 1. For better comparability, the word error rates for reconstruction based on the written (WRI) or recognized text (REC) only are given as well.

The word error rates for the written text only reconstruction are significantly higher due to formatted entities, lack of hesitation markers, and additional formatting elements like punctuations or paragraph breaks in the written texts. The hypothesized reconstruction outperformed the recognized text only reconstruction in general by about 12% relative. For texts with medium and high WER, the improvement was even higher (-14% and -15.6% relative), while for low WER, performance was slightly worse (+4.3% relative).

For the experiments based on the text length, the results did not show any specific trend. Independent from the text length, the hy-

data set	WRI	REC	HYP
all texts	34.87	22.81	20.08
$5\% \leq$ WER $\leq 13\%$	29.20	10.03	10.46
$20\% \leq$ WER $\leq 25\%$	34.97	21.04	18.09
$40\% \leq$ WER $\leq 45\%$	40.47	37.96	32.03
$0 \leq$ # words ≤ 300	34.35	21.37	18.77
$300 \leq$ # words ≤ 500	33.92	22.05	19.55
$500 \leq$ # words ≤ 1700	36.16	25.27	21.99

Table 1. Reconstruction results (WER in %)

pothesized reconstruction returned better results than the baseline methods ($\sim$12% relative improvement).

5. DISCUSSION & OUTLOOK

The results indicate that the quality of a reconstruction is directly dependent on the word error rate of the underlying recognized text. This finding is in accordance with the initial assumption that phonetic similarity matching detects phonetically similar words as they are characteristic for speech recognition errors. For lower WER, the potential for reconstruction improvement is therefore lower as well.

Many of the remaining errors still occur due to syntactic mismatches like hyphenation in concatenated words, or document formatting elements. These mismatches cannot be covered by the phonetic analysis and have to be resolved explicitly with other mechanisms in the reconstruction process.

Some errors, however, have to be attributed to a false reconstruction due to a wrong phonetic alignment. Particularly for short words like "the", "a", "and", etc., the scores returned by the similarity measure were too high, leading to a wrong reconstruction decision. An adaptive threshold for the COR/sim alignment tag could therefore also bring some improvement.

Besides, some patterns were observed in the mismatch regions that are particularly interesting for phonetic matching:

- Wrong segmentation in recognized text:
  e.g. `room ventrally` (recognized)
  `rudimentary` (written)
- Massive reductions due to fast speech:
  e.g. `stenting` (recognized)
  `understanding` (written)
- Spellings: spelled letters recognized as words
  e.g. `Aimee the ER I seek a` (recognized)
  `MAVERICK` (written)

To detect matches across word borders, it may be helpful to split concatenated words and multi-word alignment lines and then apply the scoring function again on the splitted regions. Multi-word alignment lines may not just be separated on word level, but also on syllable level to allow even slight recognition errors to be handled appropriately. This mechanism could help reduce the errors due to wrong segmentation and misrecognized spellings.

Another big improvement step should be the incorporation of semantic similarity matching in the reconstruction process [6]. This way, reformulations could be identified, formatting instructions handled more efficiently, and difficult reconstruction decisions be resolved more easily.

6. CONCLUSION

Automatic reconstruction of literal transcriptions for dictated texts is a challenging task as indicated by the description of recognized and written texts. We presented a customized phonetic similarity measure for this task which is trained on a set of phonetically similar string pairs. It returns interpretable results and is robust in its application as it includes fallback strategies to the deterministic Levenshtein measure. Furthermore, we introduced flexible automatic phonetic transcription to deal with the problem of formatted entities in written texts and alternative pronunciations in recognized texts.

In an evaluation, we showed that phonetic similarity measurement enhances the reconstruction of a literal transcription from a recognized and a written text of a dictation for recognized texts with medium to high word error rate. This finding is in accordance with the initial assumption that in general, speech recognition mismatches occur due to phonetic similarity to the actual utterance. The similarity measure still exhibits potential for refinement like better coverage of mismatches due to wrong segmentation of the recognized text or falsely recognized spellings of words.

7. ACKNOWLEDGEMENTS

We would like to thank Christina Drexel at Philips Speech Recognition Systems for providing the de-formatting grammar used in the phonetic similarity measure.

8. REFERENCES

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

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

[3] Karim Filali and Jeff Bilmes, "A dynamic Bayesian framework to model context and memory in edit distance learning: An application to pronunciation classification," in *Proceedings of the ACL*, Ann Arbor, Michigan, 2005, pp. 338–345.

[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] J. Zobel and P. W. Dart, "Phonetic string matching: Lessons from information retrieval," in *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, 1996, pp. 166–172.

[6] Martin Huber, Jeremy Jancsary, Alexandra Klein, Johannes Matiasek, and Harald Trost, "Mismatch interpretation by semantics-driven alignment," in *Proceedings of KONVENS*, Konstanz, Germany, 2006.

[7] Carnegie Mellon Univ., "The CMU pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

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

[9] Eric Sven Ristad and Peter N. Yianilos, "Learning String-Edit Distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 5, pp. 522–532, 1998.