A qualitative evaluation of phoneme-to-phoneme technology

Marijn Schraagen, Gerrit Bloothooft

Utrecht Institute of Linguistics OTS, Utrecht University, The Netherlands

{M.P.Schraagen,G.Bloothooft}@uu.nl

Abstract

Automatic speech recognition systems apply grapheme-to-phoneme transcription (G2P) to model pronunciation of items in the lexicon. General purpose G2P transcriptions are not always accurate, e.g., in a multilingual environment. To improve the transcription quality, G2P transcriptions can be postprocessed using a phoneme-to-phoneme (P2P) converter. This paper discusses the applicability of P2P technology based on results of a speech recognition experiment using P2P conversion on a multilingual speech corpus. P2P conversion can be applied successfully, however the analysis also shows limitations of P2P technology.

Index Terms: Automatic Speech Recognition, phoneme-to-phoneme conversion, multilingual speech corpus

1. Introduction

Current systems for Automatic Speech Recognition (ASR) can reach high levels of accuracy on a closed vocabulary isolated utterance task with native speakers. Performance can decrease, however, in a multilingual situation, where the native language of the speaker is different from the language of (part of) the vocabulary. In this case, both the acoustic models and the phonetic transcriptions used by the system will be less accurate for the recognition task. A native English speaker, for example, might pronounce the German word lang (English: long) as /lEnk/ while the ASR system expects the correct German pronunciation /lAnK/. A partial solution to this problem is to incorporate grapheme-to-phoneme (G2P) engines for multiple languages into the ASR system (for the example an English G2P engine is likely to generate /lEnk/). However, with multiple languages many irrelevant transcription will be produced that potentially complicate the recognition task. Moreover, a speaker in a multilingual setting is likely to have at least some familiarity to the target language, which means that the pronunciation that the speaker will use is somewhere in between the correct transcription and the transcription of a G2P for his native language. Therefore, a transformation of the correct transcription to the expected actual transcription is desired. Such a transcription is generated by a phoneme-to-phoneme converter or P2P.

1.1. Applicability of P2P technology

Transformation rules for a P2P system can be designed by hand [1], or derived automatically from training data [2, 3]. An important parameter for both approaches is the complexity of the rules. Single phoneme rules (such as /A/ → /E/) are likely to miss a significant amount of variation in non-native speech and introduce new confusion pairs for similar lexical items, while complex rules might suffer from limited applicability and overfitting on training data. A general problem in non-native pronunciation modelling is the high level of variation in pronunciation. Part of the variation is speaker dependent, which is hard to model by a P2P converter. Moreover, many variants are not systematic. If an utterance a speaker intends to say contains words that the speaker does not fully understand (which is often the case with non-native speakers), the pronunciation suffers not only from the phonetic differences between the languages but also from the lack of understanding. This type of variation is highly unpredictable.

The goal of this paper is to evaluate the applicability of phoneme-to-phoneme technology. First, an analysis is presented on the type of transformations that a P2P converter should incorporate in order to increase the accuracy of an ASR system. Second, applicability of P2P technology in general is reviewed based on observed properties of variation. The multilingual speech corpus and the P2P technology developed within the Autonomata TOO project\(^2\) are used in the analysis.

2. Methodology

The Autonomata TOO corpus is a closed vocabulary, isolated utterance speech corpus containing 80 speakers with five different native languages (Dutch, English, French, Moroccan Arabic/Berber, Turkish). All speakers were living in the Netherlands or Belgium at the time of the recording. The lexicon of the corpus consists of names of Points-of-Interest (POIs), such as restaurants and hotels, from The Netherlands and Belgium. The POIs have been selected to contain Dutch, English and French words, therefore the corpus is multilingual on both the speaker and the lexicon level. The corpus contains around 21,500 recordings. For every recording, a number of phonetic transcriptions are available: a generic G2P transcription from a Dutch, French and English G2P engine, a reference transcription for the POI by a human expert, and an auditory verified transcription of the particular recording.

The auditory verified transcription (AVT) enables an automatic analysis of the variation in the utterances. We divide the utterances into four variation categories:

1. Correct: The AVT is equal to the reference transcription.
2. Phoneme substitution: One or more phonemes from a phoneme confusion set should be substituted in the AVT to obtain the reference transcription. The phoneme confusion set is defined as all vowel phonemes and a limited number of consonant phonemes, as follows: \{A, E, S, s, u, I, O, i, o, e, y, G, g, x, f, s, V, w, A, S, u, ˘ y, e+, A˘ , E˘ , s, z, Z, j, t, d, ˘ n, nK, p, b\}

\(^2\)The Autonomata TOO corpus is available on http://www.inl.nl/en/corpora/autonomata-poi-corpus. Use of the corpus is free of charge for research purposes.
3. Structural variant: Phonemes outside of the phoneme confusion set are involved in the transformation from the reference transcription to the AVT.

4. Error: The utterance contains a severe reading error, such as a deletion, insertion or substitution of a full word.

Figure 1 shows the distribution of variation categories over speaker origin. For each speaker group, the four categories sum to 100%. The test items have been unequally distributed over the speaker groups: for the Dutch group, only 15% of all recordings consisted of an all-Dutch POI, while for all other speakers every POI is either fully in Dutch or containing both Dutch and foreign words. This explains the relatively low proportion of correct pronunciations for the Dutch speakers in Figure 1.

The Autonomata TOO toolkit provides a training algorithm for P2P converters (see [4]). Using this algorithm, P2P converters are trained to transform a G2P transcription into an expected actual transcription. The training data used for the converter are extracted from the Autonomata Spoken Name Corpus (ASNC)\(^3\). The ASNC contains G2P transcriptions, reference transcriptions and auditory verified transcriptions for a large number of recordings in the POI-related domain of person names and geographic names, for the same languages as the Autonomata TOO corpus. The resulting P2P transcriptions have been used in a speech recognition experiment on the Autonomata TOO corpus using the Nuance VoCon 3200 speech recognition engine. The results of speech recognition with P2P transcriptions are compared to the results of speech recognition using G2P transcriptions only. In general, P2P transcriptions are able to obtain a (slightly) lower error rate. P2P technology is not equally effective for every type of variation. To analyze in which situations P2P technology is most useful in improving speech recognition, the results are grouped by speaker origin and variation category (see Figure 2).

3. Effective P2P transformations

The baseline recognition error rate for the G2P on this multilingual dataset is 21.1%. The P2P converter is able to improve the recognition accuracy by 1.4% averaged over all records (6.6% relative improvement). Figure 2 shows the accuracy gain using a P2P by variation category. First, the general behaviour of the P2P converter is analyzed. The average number of transformations per transcription is 2.5, and not all transformations need to be equally effective. However, the presence of a rule in an improved transcription and the use of this rule in improving recognition accuracy are likely to be correlated. The conversion rules that contributed the most to improved recognition are listed in Table 1. Most of the rules are substitutions of similar phonemes, such as /æ/ → /ʌ/ or /ɛ/ → /ı/. The rules show that subtle differences in transcription can be sufficient to guide the ASR engine in the right direction. This also explains why even the recordings with a correct transcription can be recognized more accurately using a P2P converter. Even though a transcription is correct, in a number of cases a slightly altered transcription can be more suitable for the ASR search process.

The coverage of each rule should be considered as well: the first 15 rules (out of a total of 212) account for almost 50% of all recognition improvement. This has potential implications for the (manual or automatic) rule induction process, for example because not all phonemes need to be equally involved in rule induction. The rules in Table 1 are a simplification of the actual rules, which consider some context and additional features to prevent overgeneralization. However, the actual transformations of the transcriptions and therefore the observations on coverage are accurately described in the table.

4. Variation not captured by P2P converters

Table 2 shows transformations that were not found during P2P training. These rules are transformations from the G2P transcription towards the auditory verified transcription for incorrectly recognized recordings, which would be good candidates for P2P conversion. Only recordings from the first three categories (i.e., no severe errors) are included in the table. It should be noted that better transcriptions do not necessarily lead to improved recognition accuracy, but the transcription quality and the recognition accuracy are obviously correlated. The percentage of incorrectly recognized utterances for which a rule would improve the transcription is listed in the Coverage column. The coverage of the first 15 rules is lower than the percentage of records with improved recognition results from Table 1. This means that the missed rules have a smaller effect than the rules that are incorporated in the P2P, indicating that the transcrip-
Table 1: Phoneme transformations leading to improved recognition

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Corrections (%)</th>
<th>Cumulative (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/E/ → /A/</td>
<td>6.2 %</td>
<td>6.2 %</td>
</tr>
<tr>
<td>/a/ → /A/</td>
<td>5.0 %</td>
<td>11.2 %</td>
</tr>
<tr>
<td>/w/ → /V/</td>
<td>4.8 %</td>
<td>16.0 %</td>
</tr>
<tr>
<td>/$// → ∅</td>
<td>4.4 %</td>
<td>20.4 %</td>
</tr>
<tr>
<td>∅ → /n/</td>
<td>4.1 %</td>
<td>24.5 %</td>
</tr>
<tr>
<td>/$// → /E/</td>
<td>3.2 %</td>
<td>27.7 %</td>
</tr>
<tr>
<td>/w/ → ∅</td>
<td>2.9 %</td>
<td>30.6 %</td>
</tr>
<tr>
<td>/$// → /$/</td>
<td>2.8 %</td>
<td>33.4 %</td>
</tr>
<tr>
<td>/z/ → /s/</td>
<td>2.4 %</td>
<td>35.6 %</td>
</tr>
<tr>
<td>/e/ → /E/</td>
<td>2.4 %</td>
<td>38.0 %</td>
</tr>
<tr>
<td>/u/ → /o/</td>
<td>2.3 %</td>
<td>40.3 %</td>
</tr>
<tr>
<td>∅ → /$//</td>
<td>2.2 %</td>
<td>42.5 %</td>
</tr>
<tr>
<td>/$// → /$/</td>
<td>2.0 %</td>
<td>44.5 %</td>
</tr>
<tr>
<td>∅ → /e/</td>
<td>2.0 %</td>
<td>46.5 %</td>
</tr>
<tr>
<td>∅ → /$/</td>
<td>1.8 %</td>
<td>48.3 %</td>
</tr>
</tbody>
</table>

Table 2: Phoneme transformations (G2P → AVT) not applied by the P2P engine

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Coverage (%)</th>
<th>Cumulative (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/$// → ∅</td>
<td>5.3 %</td>
<td>5.3 %</td>
</tr>
<tr>
<td>/t/ → ∅</td>
<td>3.2 %</td>
<td>8.8 %</td>
</tr>
<tr>
<td>∅ → /A/</td>
<td>2.6 %</td>
<td>11.4 %</td>
</tr>
<tr>
<td>/z/ → ∅</td>
<td>2.5 %</td>
<td>13.9 %</td>
</tr>
<tr>
<td>/n/ → ∅</td>
<td>2.5 %</td>
<td>16.4 %</td>
</tr>
<tr>
<td>/s/ → ∅</td>
<td>2.2 %</td>
<td>18.6 %</td>
</tr>
<tr>
<td>/$// → /E/</td>
<td>2.1 %</td>
<td>20.7 %</td>
</tr>
<tr>
<td>/E/ → ∅</td>
<td>2.0 %</td>
<td>22.7 %</td>
</tr>
<tr>
<td>/a/ → ∅</td>
<td>2.0 %</td>
<td>24.7 %</td>
</tr>
<tr>
<td>/k/ → ∅</td>
<td>1.8 %</td>
<td>26.5 %</td>
</tr>
<tr>
<td>∅ → /$//</td>
<td>1.7 %</td>
<td>28.2 %</td>
</tr>
<tr>
<td>/e/ → ∅</td>
<td>1.5 %</td>
<td>29.7 %</td>
</tr>
<tr>
<td>/l/ → ∅</td>
<td>1.4 %</td>
<td>31.1 %</td>
</tr>
<tr>
<td>/a/ → ∅</td>
<td>1.3 %</td>
<td>32.4 %</td>
</tr>
<tr>
<td>∅ → /E/</td>
<td>1.2 %</td>
<td>33.6 %</td>
</tr>
</tbody>
</table>

5. Transformation position effects

The search algorithm used in the VoCon 3200 ASR engine processes the speech signal in the same order as it is recorded (i.e., the beginning of the signal is processed first). The search for
hypotheses is narrowed as the signal is being processed. This means that a lexical item that matches the initial part of the signal but not the final part has a much higher probability of being selected than a lexical item that matches the final part but not the initial part. Regarding phonetic transcriptions, this means that the beginning of the transcription has much more influence on accurate recognition than the end of the transcription. This is illustrated by Figure 3. The figure analyzes utterances for which the G2P transcription contains errors, i.e., the auditory verified transcription differs from the G2P transcription (P2P transcriptions are not used in the analysis). The position of the first error is measured, both for correctly and incorrectly recognized utterances. The figure shows the cumulative distribution of the position of errors for both correct and incorrect results. The first transcription error for incorrectly recognized utterances is located in the beginning of the transcription: 20% of the transcriptions contains the first error at the first position, 45% of the transcriptions contains the first error at position 1 or 2. For the correctly recognized utterances, the error position is more evenly distributed: 10% of the transcriptions contains an error at the first position, 30% at the first 2 positions.

The relative importance of the position of transcription variation is important for P2P research. It means that a P2P training algorithm should select rules that fit the training data at the initial part of the transcription, even if these rules do not fit the data in other transcription positions.

Another approach of improving recognition accuracy regarding error position effects could be to modify the recognition process itself. If the recognition process is able to process the speech signal in non-linear order, accuracy can improve by correcting errors at various positions in the transcription.

6. P2P user profiles

The P2P training algorithm is unable to capture low-frequent or idiosyncratic variation in the training data. In general, this is the desired behaviour, because transformation rules should generalize over speakers and utterances. However, for an individual speaker, this can result in lower recognition accuracy. If the P2P converter would be able to apply a selection of rules for a particular speaker, recognition accuracy is likely to improve. To do this, the P2P model needs to be extended to be able to generate user profiles with specific transformation rules. The concept of a user profile is known from ASR systems in general, that require a new user to record a number of training utterances. A P2P system could include a similar training phase, where various transformation rules obtained from a corpus (including low-frequent and idiosyncratic transformations) are tested on user input to determine which rules apply for the new speaker. This procedure would in effect compose a new speaker profile from parts of corpus speakers. The benefit of this approach over traditional speaker adaptation is that the derived rules can be applied to the full lexicon, instead of only the training utterances.

The use of a training phase has the disadvantage that the usability of a system in, e.g., public information services is limited. However, the analysis of the Autonomata TOO data suggests that, given the irregularity in pronunciation variation (especially for the structural variation and error categories), any general purpose P2P training phase will fail to improve recognition accuracy on a significant amount of recordings. Speaker specific solutions, such as an individual P2P training phase, provide an alternative for accuracy improvement. Other types of improvement can be obtained by using feedback strategies (see [6]), which are able to handle most utterances from the error category.

7. Conclusion

The results described above indicate that phoneme-to-phoneme technology can be successfully applied to improve recognition accuracy in ASR. The P2P converters used in the experiments already provide useful transformation rules, but the analysis of incorrectly recognized utterances shows that there is potential for expanding the rule set. The characterisation of transformation errors suggests explicit segment deletion modeling to improve transcription accuracy. However, the category analysis shows that severe errors are very hard to address using P2P technology. Other methods should be applied to improve recognition accuracy for recordings of this type.

8. References