Pronunciation-based ASR for names

Henk van den Heuvel [1], *Bert Réveil*[2], *Jean-Pierre Martens*[2]

[1] CLST, Fac. of Arts, Radboud University Nijmegen, Netherlands
[2] ELIS, Ghent University, Belgium
H.vandenHeuvel@let.ru.nl

Abstract

To improve the ASR of proper names a novel method based on the generation of pronunciation variants by means of phoneme-to-phoneme converters (P2Ps) is proposed. The aim is convert baseline transcriptions into variants that maximally resemble actual name pronunciations that were found in a training corpus. The method has to operate in a cross lingual setting with native Dutch persons speaking Dutch and foreign names, and foreign persons speaking Dutch names. The P2Ps are trained to act either on conventional G2P-transcriptions or on canonical transcriptions that were provided by a human expert. Including the variants produced by the P2Ps in the lexicon of the recognizer substantially improves the recognition accuracy for natives pronouncing foreign names, but not for the other investigated combinations.

Index Terms: ASR, name recognition, pronunciation modeling, lexicon development, multi-linguality

1. Introduction

The manual generation of phonetic transcriptions for Automatic Speech Recognition (ASR) is very time-consuming and subject to a great deal of inconsistencies. For this reason, automatic grapheme-to-phoneme converters (G2Ps) have been developed [2,4,9]. However, general purpose G2Ps often perform poorly when it comes to the transcription of names. Names typically do not adhere to the standard spelling conventions of a language due to their spellings and/or their foreign origin. Therefore, they need special treatment [3,5,12]. In practice, pronunciations of a name vary widely. This is the case for native speakers pronouncing names from their native language, but the more so in a cross-lingual context when non-native speakers pronounce these names, or when native speakers pronounce names of a foreign origin. Such cross-lingual variations are encountered at the canonical level [e.g. 10], but even more so in actual speech production.

In the Autonomata Too project we aim to improve the performance of a name recognizer by dealing more properly with the large degree of variations in name pronunciations. The objective of the present research is to improve the recognition of (1) native Dutch/Flemish pronunciations of Dutch/Flemish names, (2) native Dutch/Flemish pronunciations of foreign names and (3) non-native Dutch/Flemish pronunciations of Dutch/Flemish names. We hope to achieve that by adding effective pronunciation variants to the lexicon. In this study, we concentrate on modeling pronunciation variants found in recorded name utterances using P2Ps, whereas in a related study [6] pronunciation variants generated by non-native standard G2Ps are envisaged

In this paper we distinguish 3 kinds of (broad) phonetic transcriptions: a. G2P-transcriptions made by the Nuance standard G2P-convertors for common words, b. EXP-transcriptions being canonical name transcriptions made by human EXPerts, and c. AV-transcriptions, Auditorily Verified transcriptions, reflecting what a human expert heard when listening to the name utterance in a speech corpus. In our method self-learning P2P-convertors are used to convert initial G2P-transcriptions resp. EXP-transcriptions into variants that are closer to AV-transcriptions. The idea is that the P2P-conversion rules learned on a train set will help to bring the initial transcriptions closer to the AV-transcriptions for the test set names as well.

The method differs from that of e.g. [7] by the fact that rules are automatically learned from data rather than construed by hand; it differs from that of e.g. [1] by the fact that the variants are not learned by taking reference transcriptions that were automatically derived from the speech signals, but by using reference transcriptions that were obtained by listening to the name utterances. The training data and the P2P learning algorithm are described in more detail section 2.

In section 3, we evaluate our pronunciation variants in two ways. First we perform a transcription-based evaluation in which we compare the P2P-generated pronunciation variants with the AV-transcription of a lot of name utterances. Then we perform an ASR-based evaluation in which we compare the recognition accuracy obtained with a baseline lexicon and with a lexicon containing the P2P-generated pronunciation variants.

In the present study we will restrict ourselves to phonemes within the phoneme inventory of Dutch, and to a speech recognition engine with monolingual acoustical models only.

2. Data and set-up of the experiments

2.1. Speech data

We used the Autonomata Spoken Name Corpus (ASNC) for our experiments [11]. The corpus includes spoken utterances of 240 speakers (120 native and 120 non-native) half of which were recorded in Flanders, the other half in the Netherlands. Each speaker was asked to read 181 proper names and 50 command & control words from a computer screen. The names belonged to the categories person names (first name + family name) and geographical names (street names and city names). For this study we focus on the Flemish part of the corpus. In this part the 60 non-native speakers are evenly spread over English (EN), French (FR) and Moroccan (MR). The non-native speakers do have some language proficiency in Dutch (around A2, B1 in CEF-terms).

Table 1. *Number of name tokens per cell of the development and test set from the Flemish part of the ASNC*

Speaker origin			**#records train/dev. Corpus**			**#records test corpus**		
			Name source			**Name source**		
	NL	**EN**	**FR**	**MR**	**NL**	**EN**	**FR**	**MR**
NL	5040	966	966	630	2160	414	414	270
EN	1440	276	276	180	960	184	184	120
FR	1680	322	322	210	720	138	138	90
MR	1680	322	322	210	720	138	138	90

The material was divided in a training/development set and a test set. There was no overlap between the sets neither in terms of speakers nor in terms of names. Table 1 shows the number of name tokens per speaker tongue and name source.

The data in the shaded cells were excluded from our experiments since they fall outside the scope of our present research (and suffer from data sparseness).

2.2. Generation of pronunciation-based variants

Previously developed P2P learning tools [12] were used to generate the pronunciation-based transcription variants. They implement a four-step procedure, which is visualized in Fig. 1. First, the initial phonemic transcription is aligned with the target phonemic transcription (AV-transcriptions of the actual name pronunciations in the ASNC) and with the orthographic transcription. The second step retrieves from these alignments the input/output transformations that account for the systematic errors found in the initial transcriptions. Given these transformations, the alignments are re-used (step 3) to generate the training examples from which to learn correction rules. The final step is the actual rule induction from these examples.

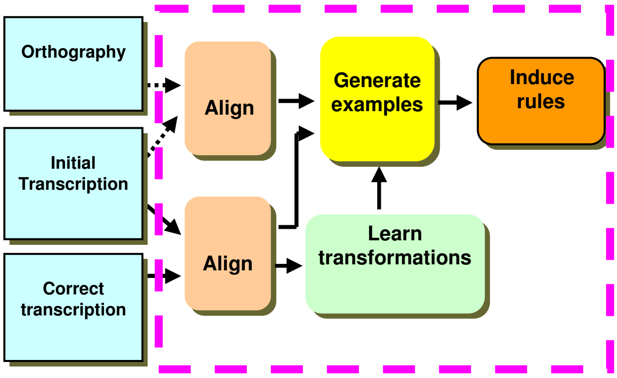

Figure 1 : *Automatic learning of the ELIS P2P converter*

In generation mode, the P2P converter runs from left to right over the aligned initial phonemic and the orthographic transcription, and it inspects both transcriptions to establish when it can apply one of its learned stochastic correction rules. Each rule expresses the following: If a particular phonemic pattern (called the *rule input*) occurs in the initial phonemic transcription and if the phonemic and orthographic context in which it occurs meets the *rule condition*, then transform the rule input to an alternative phonemic pattern (called the *rule output*) and assign a certain probability to that pattern. The rule condition can describe constraints on the identities of the phonemes at both sides of the rule input, the stress level of the syllable associated with that input, the position of this syllable in the word, etc. It can also express constraints on graphemic patterns in the orthography which is aligned to the rule input By applying rules at different places in the input transcription, the P2P converter can generate multiple pronunciations with attached probabilities.

One of the factors we investigated in this study is the origin of the initial phonemic transcription. This can be a transcription generated by the Dutch G2P of Nuance, or it can be a canonical transcription of the name as provided by a human expert (EXP-transcription). The latter is supposed to represent the most common pronunciation of the name by a native speaker. The EXP-transcriptions are generally closer to the target transcriptions than the G2P-transcriptions, and may therefore be better suited as starting points for training P2Ps. On the other hand, it is time consuming to create these transcriptions. Fortunately, TeleAtlas is a project partner and it owns such transcriptions.

Both for initial G2P-transcriptions and initial EXP-transcriptions, we trained seven P2P-converters which are trained on the development data taken from the seven non-shaded cells in Table 1. Thus, a total of 14 P2Ps was trained. Next, we let each P2P generate up to four transcriptions per name appearing in the corresponding cell of the test set in Table 1. We will call such transcriptions P2P-transcriptions. Only transcriptions with a probability that exceeds some fixed threshold of 0.2 are being generated.

In order to perform a transcription-based evaluation of the P2P performances we compared all generated P2P transcriptions of each name to the AV-transcription of that name in the test set; we selected the best of these P2P-transcriptions and we derived two measures for the discrepancy. One is the Name Error Rate (NER, which is similar to WER), defined as the number of times the best P2P-output is different from the target. The other is the relative Name Improvement Rate (rNIR), defined as the percentage of times the P2P output is better (having a lower Levenshtein distance) than its input minus the percentage of times it is worse. The decision to consider only the variant that best matches the target transcription to compute NER and rNIR is in line with the situation in ASR where one utterance is linked to the best matching variant in the lexicon.

2.3. Pronunciation-based ASR for names

For our recognition experiments we used the VOCON3200 v.3 engine provided by Nuance. For the present experiments the recognizer operates with monolingual acoustical models for Flemish Dutch only.

In experiments with native speakers, the lexicon contains all 3540 names appearing in the ASNC (train and test set, Dutch and Flemish part). This relatively large grammar was used in order not to restrict the choice of the recognizer to names in the Flemish test set (only 543 name types), which would have resulted in a relatively simple task for the baseline systems. In experiments with foreign speakers, only the Dutch/Flemish names were included in the lexicon (2339 entries), because in that case the foreign names do not constitute the target of our research. Priors of the pronunciation variants are not taken into account in the lexicon.

There were two baseline systems, one for each choice of the initial phonemic transcriptions. For the case of G2P-transcriptions, the baseline lexicon contained three G2P-transcriptions for each name, namely those which were generated by the Dutch, English and French G2P converters of Nuance and subsequently *nativized* by means of phoneme mapping. Foreign transcriptions were included because in [6] they were found to be beneficial for the recognition performance and because we wanted to evaluate the P2Ps starting from a strong baseline system. For the experiment with the P2Ps trained on EXP-transcriptions the lexicon contained the (one and only) EXP-transcription for each name.

In order to establish whether AV-transcriptions would be profitable targets during P2P-training we conducted a preliminary recognition experiment. Recognition performances obtained with baseline systems comprising G2P-transcriptions, EXP-transcriptions and AV-transcriptions were compared to one-another. The results are in Table 2. As can be seen from Table 2 there is much to gain with AV-transcriptions. Only for Flemish speakers reading Dutch/Flemish names the AV-transcriptions do not outperform the EXP-transcriptions.

Table 2. *ASR name recognition performance for G2P-transcriptions, EXP-transcriptions and AV-transcriptions, respectively. Name Error Rates in %.*

		Name source			
Speaker origin		**NL**	**EN**	**FR**	**MR**
FL	**NER(G2P)**	3.8	8.7	3.6	1.9
	NER(EXP)	2.6	6.8	4.4	4.4
	NER(AV)	2.8	2.4	3.1	0.7

		Speaker origin			
Name Source		**FL**	**EN**	**FR**	**MR**
NL	**NER(G2P)**	-	18.9	18.6	11.7
	NER(EXP)	-	19.0	19.3	11.5
	NER(AV)	-	6.3	12.9	5.1

3. Results

In this section we review the transcription based and ASR-based evaluation experiments we performed to assess the power of the P2Ps.

3.1. Transcription-based Evaluation

In the transcription-based evaluations the scoring is restricted to phoneme symbols only (syllable boundaries were kept during the alignment phase, but were removed before the scoring took place).

First, we have evaluated for the names in the test set whether the transcriptions generated by the P2Ps yield a closer match to the AV target transcriptions in the test set than the G2P-transcriptions do. A name error is found each time none of the generated P2P-transcriptions gives a perfect match. The top panel of Table 3 shows that the NER for names spoken by Flemish speakers is moderate (25.7%) in case the name source is Dutch/Flemish, but high (larger than 65%) in all other cases. Similarly, the bottom panel shows that for the pronunciation of Dutch names by different speaker categories, the NER is moderate (25.7%) for Flemish speakers and high (larger than 48%) for the non-native speakers. Nevertheless, the P2Ps cause very substantial improvements in all examined combinations of speaker and name origin (almost all rNIRs are larger than 40%). The improvements for non-native names are larger than for non-native speakers.

Table 3. *Results for the P2Ps based on G2P-transcriptions in terms of NER and relative NIR percentages.*

		Name source			
Speaker origin		**NL**	**EN**	**FR**	**MR**
FL	**NER**	25.7	65.7	76.1	73.3
	rNIR	47.5	47.3	64.8	68.7

		Speaker origin			
Name source		**FL**	**EN**	**FR**	**MR**
NL	**NER**	-	70.2	56.1	48.9
	rNIR	-	40.4	42.4	38.5

Next, the EXP-transcriptions were used as point of departure for the P2P-training. The results are in Table 4. In this case, the NER scores (see in Table 4) are lower, but this is mainly due to the fact that EXP-transcriptions much better resemble the AV-transcriptions than G2P-transcriptions. Nevertheless, the P2Ps still yield a substantial rNIR which is mostly larger than 20%. In contrast to the G2P-transcriptions, the improvements for non-native names are smaller than for non-native speakers (except for MR). This complies with the fact that the EXP-transcriptions are intended to reflect the pronunciations of native speakers so that there will be less to gain for a P2P-engine that is trained for native-names.

Table 4. *Results for the P2Ps based on EXP-transcriptions in terms of NER and relative NIR percentages.*

		Name source			
Speaker origin		**NL**	**EN**	**FR**	**MR**
FL	**NER**	15.6	44.7	44.0	63.3
	rNIR	21.7	18.3	17.7	30.3

		Speaker origin			
Name source		**FL**	**EN**	**FR**	**MR**
NL	**NER**	-	66.0	50.3	44.9
	rNIR	-	38.5	40.2	20.8

We will now investigate if the observed improvements of the transcriptions also translate into improved ASR performance.

3.2. ASR-based Evaluation

For the ASR-based evaluation, we use the NER to represent the number of names that were wrongly recognized. The rNIR in this context is defined as [NER(P2P) - NER(base)] / NER(base), where NER(base) is either the NER found for the G2P-transcriptions or for the EXP-transcriptions as baseline transcriptions .

First, we tested the P2P-transcriptions trained on the G2P-transcriptions. We obtained the best results with a lexicon that had no more than two additional P2P-transcriptions for the native speakers and three for the non-native speakers (see Table 5).

Table 5. *ASR-results for the P2Ps based on G2P-transcriptions in terms of NER and NIR percentages.*

		Name source			
Speaker origin		**NL**	**EN**	**FR**	**MR**
FL	**NER**	3.2	6.0	3.1	1.9
	rNIR	16.0	30.6	13.3	0.0

		Speaker origin			
Name source		**FL**	**EN**	**FR**	**MR**
NL	**NER**	-	17.4	17.5	11.5
	rNIR	-	7.9	5.9	1.7

Overall, the NERs are substantially lower now than those observed during the transcription based evaluation. This means that a mismatch in the transcriptions not necessarily leads to a misrecognition. Conversely, this also implies that better matching transcriptions do not necessarily lead to a better recognition performance. For native speakers the P2P-transcriptions yield a substantially improved recognition for all name categories except the Moroccan names. The largest gain is observed for the English names, where there was also most room for improvement (see Table 2).

For foreign speakers reading Dutch/Flemish names there is obviously a lot of room for improvement, too. However, the P2Ps are not capable of yielding more than a marginal gain.

Finally, we investigated the effect of P2Ps which were trained on EXP-transcriptions. Table 6 shows the corresponding ASR results for the optimal case in which no more than two P2P-transcriptions per name were added to the lexicon (both for native and non-native speakers).

Table 6. *ASR-results for the P2Ps based on EXP-transcriptions in terms of NER and NIR percentages.*

		Name source			
Speaker origin		**NL**	**EN**	**FR**	**MR**
FL	**NER**	2.9	5.8	3.9	3.3
	rNIR	-10.6	14.2	11.3	25.0

		Speaker origin			
Name source		**FL**	**EN**	**FR**	**MR**
NL	**NER**	-	19.1	18.2	11.0
	rNIR	-	-0.5	5.7	4.3

We observe the same pattern as for the P2Ps trained on G2P-transcriptions. Substantial improvements are found for the Flemish speakers of foreign names (especially of English names) but not for the foreign speakers of Dutch/Flemish names. There is even a marginal deterioration for Flemish speakers of Dutch Flemish names. It seems that the baseline system already reached a ceiling performance for this category (cf. the corresponding NER (EXP) and NER(AV) scores in Table 2).

From comparing the NER-scores in Tables 5 and 6, we can conclude that EXP-transcriptions do not lead to better results than G2P-transcriptions when used in combination with automatically trained P2Ps.

4. Discussion

The results in section 3.1 show that the P2Ps trained on G2P-transcriptions and on canonical transcriptions can pick up a great deal of regularities present in the discrepancies between the G2P-transcriptions and the AV-transcriptions. High rNIR scores for transcription comparisons attest this. This holds both for native and non-native speakers and for all name sources. Furthermore the NER scores show that the EXP-transcriptions are closer to the AV-transcriptions than the G2P-transcriptions, which is not a surprise taking into account that G2Ps are designed for common words in Dutch, not for names.

The gains observed at the transcription level only partly translate into improved automatic name recognition. Improvement is indeed achieved for foreign names spoken by native speakers, but for foreign speakers the improvement is marginal to nil. Despite the fact that the P2P's improve the transcriptions, they do not bridge the gap to the real pronunciations.

Furthermore, when used in combination with an automatically learned P2P, the EXP-transcriptions do not show better recognition results than the G2P transcriptions do. The good news about that is that, once one has a good G2P, one does not need the relatively costly EXP-transcriptions to obtain good results with our method.

Our results were obtained under the assumption that the origin of the names and the speakers is known beforehand. In making applications one can know the language origin of (most of) the names in advance, but not the origin of the speakers. One could envisage language-specific applications, if the differences between speakers of different language origins are substantial. However, the results obtained with our approach so far cast doubt on the idea that such a distinction in mother tongue of non-native speakers is effective for the recognition of native names.

In our future work we will investigate whether we can find a better selection of pronunciation variants, e.g. by combining pronunciation variants from P2Ps trained on G2P-transcriptions and EXP-transcriptions. Further, an extension to the case of a multilingual phoneme inventory and multilingual acoustic models is in progress. In a next step, we will extend the method to the broader class of Points of Interest (POIs).

5. Acknowledgements

The presented work was carried out in the Autonomata Too project, granted under the Dutch-Flemish STEVIN[1] program. The project partners are the universities of Ghent, Nijmegen and Utrecht and the companies Nuance and TeleAtlas.

6. References

[1] Beaufays, F., Sankar, A., Williams, S., Weintraub, M. (2003) “Learning linguistically valid pronunciations from acoustic data.” Proceedings Eurospeech 2003, Geneva, 2593-2596.

[2] Bisani M., Ney H. (2003) “Multigram-Based Grapheme-to-Phoneme Conversion for LVCSR”, Procs. Interspeech, 933-936.

[3] Boula de Mareüil, P.; d’Alessandro, C.: Bailly, G.; Béchet, F.; Garcia, M.; Morel, M.; Prudon, R. and Véronis, J (2005). “Evaluating the pronunciation of proper names by four French grapheme-to-phoneme converters”, Procs. Interspeech, Lisbon, 1521-1524.

[4] Bouma, G. (2000). “A finite state and data-oriented method for grapheme to phoneme conversion”, Procs. ACL, 303-310

[5] Font Llitjós, A. and Black, A.W., “Evaluation and collection of proper name pronunciations online”, Proc. LREC2002, Gran Canaria, 2002, 247-254.

[6] Réveil, B., Martens, J.P., D’hoore, B. (2009) “How speaker tongue and name source language affect the automatic recognition of spoken names", These proceedings.

[7] Schaden, S. (2006): “Regelbasierte Modellierung fremdsprachlich akzentbehafteter Aussprachevarianten.” PhD Dissertation University of Duisburg.

[8] Stouten, F., Martens, J-P. (2007). “Recognition of foreign names spoken by native speakers”, Procs. Interspeech, 2133-2136.

[9] Taylor, P. (2005), “Hidden Markov Models for grapheme to phoneme conversion”, Procs Interspeech 2005, Lisbon, 1973-1976.

[10] The Onomastica Consortium (1995). “The ONOMASTICA interlanguage pronunciation lexicon.”, Proceedings EUROSPEECH-95, Madrid, Spain, 829-832.

[11] Van den Heuvel, H., Martens, J-P., D’hoore, B., D’Haene, C., Konings, N. (2008) “The Autonomata Spoken Names Corpus” , Procs. LREC, Marrakech

[12] Yang, Q., Martens, J.P., Konings, N., Van den Heuvel, H. (2006). “Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names", Procs. LREC, Genua, 287-292.

[1] http://taalunieversum.org/taal/technologie/stevin/