

How speaker tongue and name source language affect the automatic recognition of spoken names

Bert Réveil¹, Jean-Pierre Martens¹, Bart D'hoore²

¹DSSP, ELIS, Ghent University, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium

²Nuance, Guldensporenpark 32, 9820 Merelbeke, Belgium

bert.reveil@elis.ugent.be, jean-pierre.martens@elis.ugent.be, bart.dhoore@nuance.com

Abstract

In this paper the automatic recognition of person names and geographical names uttered by native and non-native speakers is examined in an experimental set-up. The major aim was to raise our understanding of how well and under which circumstances previously proposed methods of multilingual pronunciation modeling and multilingual acoustic modeling contribute to a better name recognition in a cross-lingual context. To come to a meaningful interpretation of results we have categorized each language according to the amount of exposure a native speaker is expected to have had to this language. After having interpreted our results we have also tried to find an answer to the question of how much further improvement one might be able to attain with a more advanced pronunciation modeling technique which we plan to develop.

Index Terms: speech recognition, proper names, pronunciation modeling

1. Introduction

The automatic recognition of proper names in a car navigation or a directory assistance application is still a challenge. First of all, the involved names often exhibit archaic spellings and originate in part from foreign languages. Furthermore, non-native users have to be accommodated. Consequently, transcriptions emerging from a native grapheme-to-phoneme (g2p) converter cannot capture the large variety of pronunciations to deal with [1]. In [2] and [3] one therefore includes transcriptions for proper names (first names and family names) computed by foreign g2p converters. To comply with the monolingual acoustic models, the g2p outputs were *nativized* by mapping each foreign phoneme to its closest native equivalent. In [2] English and French transcriptions were included in a pronunciation dictionary containing Dutch, English, French and other names. 75% of the test set speakers were Dutch natives and 25% were foreign speakers (English, French and other). The name error rate (NER) dropped by about 40% for native Dutch speakers, 70% for French speakers, 45% for English speakers and over 10% for the other foreign speakers. However, the vocabulary size was small (< 500 entries) and therefore not representative of a car navigation task for instance. In [3], eight g2p converters (for Mandarin Chinese, Czech, French, German, Hindi, Italian, Russian and Spanish) were available for creating extra foreign transcriptions for 24K proper names of various origins. A 25% reduction of the NER for foreign names spoken by speakers of the name source language was achieved, as well as a 10% reduction for foreign names spoken by US natives.

In [4, 5, 6] a set of rules is applied to create accented pronunciations. In [4], they were learned from alignments of the

native (German) transcription with the outputs of a non-native (English) phoneme recognizer. In [5] and [6] manually compiled phonological rules were employed. The drops in NER were mostly moderate (5 to +10%) with some rare degradations as well, but for English, Italian, Portuguese, Spanish and Turkish speakers reading French names, improvements of around 15% were found.

Many authors [5, 7, 8] argued that including better phonetic transcriptions in the lexicon is not enough. One also needs acoustic models that can represent accented sounds that rarely occur in regular native training material.

In [5] French acoustic models were further trained on American English, German and Spanish speech and their nativized transcriptions. In combination with the previously mentioned phonological rules the recognition of French words and expressions spoken by American English, German and Spanish speakers improved a lot: a 25% reduction of the error rate. There was also a reduction of 12.5% for the non-native speakers with other mother tongues.

Other authors [7, 8] do not nativize the foreign transcriptions and include multilingual acoustic models trained on multilingual training data. In [8], the native language is German, and the foreign training material consists of English speech spoken by Germans. In a German spoken dialogue system that answers questions about (English) movies, theatres, timetables, etc., the NER dropped by 25% over that of a monolingual system with a knowledge-based mapping of English to German (native) phonemes.

In this paper we mainly investigate the most successful approaches as a function of name source (language of origin of the name) and speaker tongue (mother tongue of the speaker). We produce some new results which have raised our understanding of the mechanisms that are responsible for producing the observed recognition improvements. At the end of the paper however, we also wonder how much our best system until now could be further improved by means of more advanced pronunciation modeling.

2. Experimental set-up

For our experimental study we needed a spoken name database, g2p converters for several languages and a speech recognizer.

2.1. Spoken name database

Since we aim to investigate name source and speaker tongue as dependent variables, we opted for a corpus that is balanced with respect to these variables. The Autonomata Spoken Name Corpus (ASNC)[9] contains utterances of Dutch, English, French, Moroccan and Turkish person names (first name + family name)

and geographical names (street names and city names) spoken by 120 Dutch, 40 English, 20 French, 40 Moroccan and 20 Turkish speakers. Each speaker has read 181 names: (1) 120 Dutch names (40 person names and 80 geographical names), (2) 23 English names (7 person names and 16 geographical names), (3) 15 Moroccan person names and (4) either 23 French names (in Flanders) or 23 Turkish names (in the Netherlands) (7 person names and 16 geographical names). The reason for the latter distinction is that many speakers in Flanders are familiar with French (the second language in Belgium) whereas speakers in the Netherlands are not.

There were 20 different name lists: 12 were read by 16 speakers, the other 8 by only 6. Because of a few overlaps between the lists that were used in Flanders and in the Netherlands, there were only 3540 different names, rather than the expected $20 \times 181 = 3620$.

For experimentation, the corpus was divided in a train set and a test set, and there was no overlap in speakers nor name lists between both parts.

2.2. Phonetic transcriptions

For generating phonetic transcriptions we disposed of the Dutch, English, French and German g2p converters that are embedded in the Nuance RealSpeak text-to-speech systems for the respective languages¹. The German g2p converter was also included because many Dutch speakers are familiar with German, and because we will test a multilingual recognizer which has seen German speech during training as well.

From the ASNC we also dispose of auditorily verified transcriptions (1 per name utterance). The auditorily verified transcription of an utterance is the best nativized transliteration of what a human expert actually heard after listening.

2.3. Recognition system

The experiments were executed with the commercially available Nuance VoCon 3200 recognizer¹. In order to investigate the effect of moving from monolingual to multilingual acoustic models, the engine came with two sets of acoustic models:

- AC-MONO: the standard VoCon 3200 Dutch models, trained on speech of native Dutch speakers from the Netherlands and Belgium. The underlying phoneme set consists of 45 phonemes, and the size of the model is 2.7 MB.
- AC-MULTI: a multilingual acoustic model, trained on a data pool consisting of the same set as AC-MONO, enhanced with UK English, French and German speech. The Dutch portion comprises about 20% of the total amount of training data. The underlying phoneme set consists of 80 phonemes. The size of the model is 4.5 MB and it contains roughly 70% more parameters than AC-MONO. Models for phonemes common to more than one language have seen data from those languages.

As a grammar we chose a name list in which all the 3540 names are placed in parallel. As a performance measure we will adopt the name error rate (NER).

3. Experimental study

The basic hypothesis underlying our experimental study is that the explanation of results is facilitated by introducing three language categories: (1) the native language (NAT) which is the

¹<http://www.nuance.com/>

target language of the application, (2) non-native languages many native speakers may have been exposed to (NN1), and (3) other non-native languages (NN2). In our case, Dutch is the native language, English and French are NN1 languages and Turkish and Moroccan are NN2 languages. The language categorization is applied on the name source as well as on the speaker tongue. Table 1 shows the emerging division of the test set.

Table 1: Number of name utterances in the test set for the different name sources and speaker tongues

speaker tongue	Name source			
	NAT	NN1	NN2	All
NAT	4400	1265	992	6697
NN1	2520	805	476	3801
NN2	2280	575	584	3439
All	9200	2645	2052	13937

3.1. Baseline system

Our baseline system uses monolingual acoustic models (AC-MONO) and Dutch g2p transcriptions. The NERs of this system are listed in Table 2. As expected, the recognition of native

Table 2: NER (%): AC-MONO / native g2p

Speaker tongue	Name source			
	NAT	NN1	NN2	All
NAT	3.9	22.5	12.6	8.7
NN1	18.1	37.5	14.7	21.8
NN2	22.5	36.4	29.3	26.0
All	12.4	30.1	17.8	16.6

names by native speakers is already quite reliable but as soon as cross-lingual effects come into play, the NERs are substantially higher. Note that for all speaker categories the recognition of NN2 names is substantially better than that of NN1 names. We tested two hypotheses in this respect.

Our first hypothesis that speakers use native g2p-knowledge to read these names was not supported by the data. The mean discrepancy (in phonetic symbol difference rate) between native g2p transcriptions and auditorily verified transcriptions appeared to be the same for NN1 and NN2 names.

A second hypothesis was that NN2 names are easier to recognize because they bear less affiliation with the native language. This hypothesis is confirmed by the fact that 60% of the misrecognized NN1 names are confused with a Dutch name, while only 40% of the NN2 names were.

In the subsequent experiments we will assess how the NERs are affected by changes in the system configuration. For each experiment the results are summarized in a table which shows the obtained NERs and which encodes the substantial gains (in bold) or losses (in italic) with respect to some reference system which is mentioned between brackets in the table caption. A gain/loss is called substantial if it is larger than 20%.

3.2. Experiment 1: adding nativized foreign transcriptions

In a first experiment, the Dutch g2p transcription of a name was supplemented with nativized French and English (NN1 languages) g2p transcriptions. The utilized phoneme mappings were suggested by a human expert.

Table 3 shows that the inclusion of foreign transcriptions helps a lot to improve the recognition of NN1 names, especially when uttered by native speakers (error rate is almost divided by three). This differs from [2, 3] where the largest improvements were found for NN1 speakers uttering NN1 names. In about 95% of the cases where an improvement for English (French) names was found, the English (French) transcription was chosen. Our results thus support the idea advocated in [10] that native speakers use their NN1 language knowledge for reading NN1 names. Most NN2 speakers also seem to use their knowledge of English (a world language) or French (second official language of Morocco).

Table 3: *NER (%)*: AC-MONO / NAT + nativized NN1 g2p's (reference = baseline system)

Speaker tongue	Name source			
	NAT	NN1	NN2	All
NAT	4.0	8.5	11.3	5.9
NN1	16.6	21.1	10.9	16.8
NN2	22.9	28.4	28.1	24.7
All	12.1	16.7	16.0	13.5

The substantial NER gain for NN1 speakers reading NN2 names is owed to the fact that Moroccan names had a French spelling, thus calling for a French-like pronunciation. Finally, it is noted that the presence of foreign transcriptions does not hurt the recognition of native names by native speakers.

3.3. Experiment 2: multilingual acoustic models

Moving to multilingual acoustic models in combination with plain NN1 transcriptions (not nativized) leads to substantial improvements for NN1 speakers (see Table 4). As argued in [10],

Table 4: *NER (%)*: AC-MULTI / NAT + NN1 g2p's (reference = experiment 1)

Speaker categories	Name categories			
	NAT	NN1	NN2	All
NAT	4.9	5.9	8.9	5.7
NN1	11.6	7.1	6.9	10.0
NN2	21.6	21.0	20.0	21.3
All	10.8	9.6	11.6	10.7

NN1 speakers can have a strong accent and produce sounds of their mother tongue. The recognition of these speakers thus improves by the availability of appropriate acoustic models for these sounds. This is most strong for NN1 names incorporating more of these sounds. For these names the NER drops from 21.1 to 7.1%.

Two other results are that both NN1 and NN2 names benefit from the multilingual acoustic model set (as in [5]) and that there is a substantial increase of the NER for native speakers reading native names. We owe the latter to the fact that a large part of the phonemes appear in different languages, and consequently, the acoustic models for these phonemes are now in part 'contaminated' by foreign pronunciations. In order to explain the first result we have conducted a control experiment in which the multilingual transcriptions were replaced by the nativized transcriptions of experiment 1. In that case the recognizer cannot call its foreign phoneme acoustic models anymore.

Consistent with the above explanation we observe an increase of the NER from 7.1 to 9.8% for the (NAT,NAT) combination. However, more surprisingly, there were no significant differences for the other combinations. Apparently, the improvement for NN1 and NN2 names across speaker tongues mainly stems from the fact that multilingual acoustic models of native phonemes are now less specialized in the recognition of native sounds, and therefore beneficial to the recognition of accented sounds.

3.4. Experiment 3: adding German transcriptions

Since the acoustic training data also comprised German speech, and since German can be considered as a NN1 language, we have also investigated the effect of adding a German g2p transcription per name to the lexicon of experiment 2. Table 5 shows that the recognition of native names spoken by native

Table 5: *NER (%)*: AC-MULTI / 4 g2p's (reference = experiment 2)

Speaker tongue	Name source			
	NAT	NN1	NN2	All
NAT	5.2	5.8	6.0	5.4
NN1	10.5	7.1	5.9	9.2
NN2	20.6	21.2	17.3	20.3
All	10.4	9.5	9.2	10.1

speakers further degrades, but that the recognition of NN2 names by native speakers improves. An analysis of the previously incorrect but now correctly recognized NN2 names learned that most of them contain one or more occurrences of the character 'u' (Curukluk Sokagi, Butrus Benhida, Urmond, Oglumus Rasuli,...). A Dutch g2p converter converts this character to /ʌ/ (like in *mud*) or /y/ (no English equivalent, like in the French *écru*) whereas a German g2p converter will more likely return a /u:/ (like in *boot*) or /U/ (like in *book*). Apparently, many Dutch speakers pronounce these names with one of the latter phonemes.

3.5. Experiment 4: using the name origin

If the name source were known, one could be more selective in adding foreign transcriptions, in the hope of eliminating the loss of accuracy for native names spoken by native speakers. Therefore we set up a test with Dutch transcriptions for Dutch names, Dutch and English transcriptions for English names, Dutch and French transcriptions for French names and all four transcriptions for NN2 names. Table 6 confirms our expectation, but it also shows that the recognition of native names by NN1 spea-

Table 6: *NER (%)*: AC-MULTI / name source specific phonetic transcriptions (reference = experiment 3)

Speaker tongue	Name source			
	NAT	NN1	NN2	All
NAT	4.2	5.3	5.3	4.9
NN1	13.7	7.6	5.0	11.3
NN2	21.5	20.4	14.2	20.1
All	11.3	9.3	7.8	10.4

kers substantially degrades. These speakers use their English / French knowledge to pronounce Dutch names.

3.6. Experiment 5: using the speaker origin

We also performed an experiment in which the transcription selection was guided by the speaker tongue: Dutch transcriptions for Dutch speakers, Dutch and English transcriptions for English speakers, Dutch and French transcriptions for French speakers and all four transcriptions for Moroccan and Turkish speakers. As anticipated, the recognition of foreign names spoken by native speakers was degraded. It went back to the level of the baseline system. The other combinations were less or not affected.

4. Pronunciation modeling

The former experimental study revealed what can be achieved with a pronunciation model based on existing general-purpose g2p converters. In this section we briefly explore what further improvements might be possible with a more advanced pronunciation modeling approach.

Imagine that one could construct a transcription generator that can convert a baseline transcription of a name into a number of transcriptions which always comprises the 'true' transcription of that name. How good would the recognition then be? We tested the case where the lexicon contained the four g2p transcriptions of experiment 3, plus all the nativized auditorily verified transcriptions found in the ASNC (this yielded 11.3 transcriptions per name on average). The noted improvements (see Table 7) were substantial for all speaker tongue and name source combinations. It comes as no surprise that the largest gains can be obtained for NN2 speakers reading NN2 names. In

Table 7: NER (%): AC-MULTI / 4 g2p's + all auditory verified transcriptions for each name (reference = experiment 3)

Speaker origin	Name source			
	NAT	NN1	NN2	All
NAT	3.7	2.6	1.7	3.2
NN1	5.3	3.7	1.9	4.6
NN2	10.5	6.3	4.8	8.8
All	5.8	3.7	2.6	4.9

spite of the utopic transcription generator that was used, the experiment does indicate that phonological transformation rules which aim to convert g2p transcriptions into 'true' transcriptions may produce a lexicon with a large potential. First attempts in this direction already yielded promising results which are described in another paper that was submitted to this conference [11].

5. Conclusions and future work

We have carefully analyzed the impact of pronunciation modeling and acoustic modeling approaches on the performance of automatic name recognition as a function of two dependent variables: the mother tongue of the speaker (the speaker tongue) and the language of origin of the name (the name source). Both variables were investigated according to three language categories: (1) native, (2) non-native but to some extent familiar to many native speakers, and (3) non-native and not familiar to native speakers.

Just adding transcriptions of foreign g2p converters (after mapping the foreign phonemes to native phonemes) leads to substantial gains for names whose source is covered by one of the added g2p converters. The gains comply with the ability of the

speakers to use their knowledge of the name source language when reading a name.

Introducing acoustic models trained on multilingual speech data showed additional gains for all non-native name categories, be at the expense of a substantial loss in the recognition of native names uttered by native speakers. We found evidence that most of the improvement actually stems from the fact that the acoustic models are less specialized in representing the native sounds, and therefore beneficial to the recognition of any kind of non-native sounds. Moreover, if the name source is known, the loss of accuracy in the case of native speakers reading native names can be eliminated by not including foreign transcriptions for native names.

Finally, we demonstrated that substantial further improvements may be possible with a more advanced pronunciation modeling approach that can generate a set of transcriptions comprising one that is close to the 'true' transcription of the name.

6. Acknowledgements

The presented work was carried out in the context of the Autonomata Too research project, granted under the Dutch-Flemish STEVIN program.

7. References

- [1] Yvon, F.; Boula de Mareuil, P.; dAlessandro, C.; Aubergé, V.; Bagein, M.; Bailly, G.; Béchet, F.; Foukia, S.; Goldman, J.; Keller, E.; OShaughnessy, D.; Pagel, V.; Sannier, F.; Véronis, J and Zellner, B. (1998). "Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French", in Computer Speech and Language 12, 393-410.
- [2] Cremelie, N. and ten Bosch, L. (2001). "Improving the recognition of foreign names and non-native speech by combining multiple grapheme-to-phoneme converters", in Proc. ISCA ITRW on Adaptation Methods for Speech Recognition, 151-154, Sophia Antipolis, France.
- [3] Maison, B.; Chen, S. and Cohen, P. (2003). "Pronunciation modeling for names of foreign origin", in Proc. ASRU, 429-434, Virgin Islands, USA.
- [4] Goronzy, S.; Rapp, S. and Kompe, R. (2004). "Generating non-native pronunciation variants for lexicon adaptation", in Speech Communication 42, 109-123.
- [5] Bartkova, K. and Juvet, D. (2007). "On using units trained on foreign data for improved multiple accent speech recognition", in Speech Communication 49, 836-846.
- [6] Bonaventura, P.; Gallochio, F.; Mari, J. and Micca, G. (1998). "Speech recognition methods for non-native pronunciation variants", in Proc. ESCA Workshop on Modeling Pronunciation Variation for ASR, 17-22, Rolduc, The Netherlands.
- [7] Van Leeuwen, D. and Orr, R. (1999). "Speech recognition of non-native speech using native and non-native acoustic models", in Proc. workshop MIST, Leusden, The Netherlands.
- [8] Stemmer, G.; Nöth, E. and Niemann, H. (2001). "Acoustic modeling of foreign words in a German speech recognition system", in Proc. Eurospeech, 2745-2748, Aalborg, Denmark.
- [9] Van den Heuvel, H.; Martens, J.-P.; D'hoore, B.; D'Haene, C. and Konings, N. (2008) "The Autonomata Spoken Names Corpus", in Proc. LREC, Marrakech.
- [10] Stouten, F. and Martens, J.-P. (2007). "Recognition of foreign names spoken by native speakers", in Proc. Interspeech, 2133-2136, Antwerp, Belgium.
- [11] Van den Heuvel, H.; Réveil, B. and Martens, J.-P. (2009), "Pronunciation-based ASR for names", submitted to this conference.