Improving proper name recognition by means of automatically learned pronunciation variants

Bert Réveil[a,*], Jean-Pierre Martens[a], Henk van den Heuvel[b]

[a] *DSSP group, ELIS, UGent, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium*
[b] *CLST, Fac. of Arts, Radboud Universiteit Nijmegen, The Netherlands*

Abstract

This paper introduces a novel lexical modeling approach that aims to improve large vocabulary proper name recognition for native and non-native speakers. The method uses one or more so-called phoneme-to-phoneme (P2P) converters to add useful pronunciation variants to a baseline lexicon. Each P2P converter is a stochastic automaton that applies context-dependent transformation rules to a baseline transcription that is generated by a standard grapheme-to-phoneme (G2P) converter. The paper focuses on the inclusion of different types of features to describe the rule context – ranging from the identities of neighboring phonemes to morphological and even semantic features such as the language of origin of the name – and on the development and assessment of methods that can cope with cross-lingual issues. Another aim is to ensure that the proposed solutions are applicable to new names (not seen during system development) and useful in the hands of product developers with good knowledge of their application domain but little expertise in automatic speech recognition (ASR) and speech corpus acquisition. The proposed method was evaluated on person name and geographical name recognition, two economically interesting domains in which non-native speakers as well as non-native names occur very frequently. For the recognition experiments a state-of-the-art commercial ASR engine was employed. The experimental results demonstrate that significant improvements of the recognition accuracy can be achieved: large gains (up to 40% relative) in case prior knowledge of the speaker tongue and the name origin is available, and still significant gains in case no such prior information is available.

Keywords: proper name recognition, pronunciation variation modeling, cross-linguality

1. Introduction

Proper name recognition is one of the automatic speech recognition (ASR) tasks that is most widely put to practice. Contemporary examples of applications that use name recognition as a key component are call routing services, where one relies on the correct understanding of the recipient's name; automated travel assistance applications that need precise knowledge of the desired destination as well as of the booker's personal data; and voice-driven navigation systems (GPS) that depend on a flawless decoding of a city name, a street name or a Point of Interest (POI) (e.g. the name of a restaurant, a museum, etc.).

Despite the fact that proper name recognition is already successfully used in various commercial applications, several research groups are engaged in trying to improve it further. Name recognition is particularly challenging because of the mismatch that often exists between the way names are represented in the recognition system (by means of phonemic transcriptions and acoustic models) and the way they are actually pronounced by the user of the system. We distinguish three (interacting) causes for this mismatch.

A first cause is the possible lack of a plausible name transcription for some of the names. When manual transcriptions are too costly to collect, the phonemic transcriptions are generated by a grapheme-to-phoneme (G2P) converter. However, this converter is usually trained on common text material and therefore not well prepared to deal with archaic name spellings and name parts originating from a foreign language. As an example we refer to the Dutch town name "Schaijk" and the mixed French-Dutch street name "Haînautlaan". While the most plausible pronunciations for these names are

*Telephone: +32 (0)9 264 33 97, Fax: +32 (0)9 264 35 94
Email addresses: breveil@elis.ugent.be (Bert Réveil), martens@elis.ugent.be (Jean-Pierre Martens), h.vandenheuvel@let.ru.nl (Henk van den Heuvel)

/“sxa:jk/[1] and /E:.“no:.la:n/ respectively, the transcriptions generated by a Dutch G2P are /sxa:.“Eik/ and /he:.“nAut.la:n/.

A second cause is that multiple and very different pronunciations exist for some of the names. This implies that even though a correct name transcription is present in the lexicon, the recognition might still fail when an alternative pronunciation is used instead. The Dutch first name “Hadewych” for instance has four plausible pronunciations: /“ha:.d@.wEix/, /“ha:.d@.wEik/, /“ha:.d@.wik/ and /“ha:.d@.wix/. The first name “Roger” has an English (/“rQ.dZ@/) and a French (/RO.“Ze/) pronunciation.

A third cause is associated with the language proficiencies of the user. A speaker may either use his foreign G2P knowledge to pronounce a foreign name, or he may stick to his native language G2P knowledge. Similarly, a speaker may articulate a foreign sound in an accented way (Van Compernolle, 2001). Both mechanisms can be responsible for a lot of different acoustic realizations of the same name. E.g., Dutch native utterances of the English street name “Milkwood Road” were found to range from a Dutch (/mIl.“kwo:t#“ro:t/) over an accented English (/“mIl.kwut#“ro:t/) to a close-to-native English pronunciation (/“mIl.kwUd#“rowd/). The amount of acoustic variation increases even further when non-native speakers from different origins come into play.

Pronunciation variation is no exclusivity of proper names, and therefore it has been investigated for a long time in the context of large vocabulary continuous speech recognition (LVCSR) as well (see (Strik and Cucchiarini, 1999; PMLA, 2002) for a survey). Most of the proposed approaches can be classified as either *lexical modeling* or *acoustic modeling/adaptation* approaches. Lexical modeling deals with variations in phonetization (which sounds did the speaker intend to utter?). It tries to add plausible alternative pronunciations of each lexical entry to a baseline pronunciation lexicon, usually comprising only one or a very few pronunciations per entry. Acoustic modeling deals with variations in the articulation of the intended sounds. By optimizing the acoustic model parameters one tries to cover all common native and/or non-native speech sound articulations. Whereas acoustic modeling was successful in adapting to several speech styles, lexical modeling only helped when used in combination with context independent acoustic models or for non-native speech recognition. In fact, (Adda-Decker and Lamel, 1999) noticed that, for both read and spontaneous native speech recognition, better acoustic models seem to demand less pronunciation variants incorporated in the lexicon. Additional pronunciation variants can even maliciously raise the lexical confusability, and lead to more recognition errors (Fosler-Lussier et al., 2005).

In the particular case of proper name recognition however, the chance that the lexicon does not hold an observed phonetization is presumably much larger than in common speech recognition, and therefore, we argue that it will benefit more from lexical modeling. In this work, we explore ways to automatically derive useful pronunciation variants for names using a limited set of training examples. We believe that such an approach can substantially reduce the manual labour of lexicon developers.

The outline of this paper is as follows. First we review some formerly proposed pronunciation variation modeling techniques. Then, in sections 3, 4 and 5 we subsequently present the newly proposed methodology, the experimental framework and the results we obtained using this framework. The last section summarizes the most relevant results and proposes ideas for future work.

2. Pronunciation variation modeling

The work of (Jurafsky et al., 2001) revealed that acoustic triphone modeling can also be employed to capture phone substitutions and deletions. However, it can not properly model more substantial variations causing e.g. syllable deletions. Apparently, both acoustic and lexical modeling can deal with phonetization variation to some extent, but nevertheless, they are expected to have complementary strengths and weaknesses.

2.1. Acoustic modeling

It has been demonstrated (Lawson et al., 2003) that acoustic models trained on native speech perform poorly on accented speech of non-natives. A well proven recipe to improve their performance is to adapt the native models to accented speech using adaptation methods such as maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995) and maximum a posteriori (MAP) adaptation (Gauvain and Lee, 1994). In (Mayfield-Tomokiyo and Waibel, 2001), this technique yielded a 25% improvement for the recognition of English text spoken by Japanese natives with a low-proficiency in English. In (Bouselmi et al., 2006),

[1] All phonemic transcriptions in this paper follow the SAMPA notation for Dutch (http://www.phon.ucl.ac.uk/home/sampa/dutch.htm), but extended with /“/, /./ and /#/ to indicate a primary stress, a syllable bound and a word bound respectively.

MLLR and MAP adaptation were used sequentially to adapt context-independent native acoustic models to an a priori known accent. They found improvements of over 50% in the context of an automated vocal command system.

An alternative approach is to consider a multilingual phoneme set and multilingual training data, and to train phonemes on data from all languages in which they appear. By doing so for a bilingual set-up (German as the native and English as the foreign language), (Stemmer et al., 2001) could improve the recognition of (partly) English movie titles uttered by German natives by 25% relative.

In (Bartkova and Jouvet, 2006, 2007) the more challenging case of multiple foreign accents was considered. French commands and expressions uttered by speakers from 24 different countries were recognized using a baseline French system, and with two so-called multilingual systems. The first multilingual system was obtained by adding three foreign (English, German and Spanish) acoustic model sets trained on speech data from the corresponding language. The second one was obtained by adding three adapted French models in which French phonemes that also occur in the foreign languages were adapted on the basis of speech data from that foreign language. In both cases the foreign models were put in parallel with the baseline French model. The multilingual acoustic models of the first type did improve the recognition for English and Spanish speakers (by about 15% to 20%), but not for the other speakers, including the German speakers (the degradation was about 25% for the French and German speakers, and around 20% on average for the other speakers). The multilingual models of the second type usually degraded the performance.

The above discussion seems to suggest that training the acoustic models with multilingual speech data or adapting native acoustic models to accented speech will lead to better cross-lingual proper name recognition. However, this can be at the expense of a degraded recognition accuracy for native and non-modeled non-native proper name utterances.

2.2. *Lexical modeling*

The aim of lexical modeling is to create a pronunciation dictionary that comprises all needed phonetizations to achieve a good recognition in combination with the available acoustic models. To that end, lexical modeling usually supplements a baseline pronunciation dictionary with extra phonetizations that might be encountered during operation of the recognizer.

Baseline phonemic transcriptions (base forms) are often in part retrieved from an electronic phonetic dictionary (e.g. (CMU, 2010)) containing one or more manually checked typical (TY) transcriptions per entry. However, such dictionaries may not contain all the words of the desired vocabulary (e.g. the proper names), and in most cases, the transcriptions of these missing words are then created by means of a general-purpose grapheme-to-phoneme (G2P) converter.

Since most of the work on lexical modeling has been conducted in the context of large vocabulary continuous speech recognition (LVCSR) and since the more specific work on proper name recognition is often inspired by this work, we divide our review of the literature into a general and a specific part.

2.2.1. *Lexical modeling for LVCSR*

A popular approach to lexical modeling is to apply transformation rules on the base form in order to create alternative transcriptions. In most cases, these rules express that a single phoneme in the base form can be transformed to an alternative phoneme (or a place holder/empty symbol) in a particular context. The context is usually restricted to the identities of the phonemes immediately to the left and to the right of the examined phoneme.

The transformation rules can either be created manually by a linguist, like in e.g. (Bonaventura et al., 1998; Wester et al., 2000; Schaden, 2003a,b; Bartkova and Jouvet, 2007), or they can be learned automatically from correspondences between base forms and observed pronunciations. This so-called data-driven approach is exemplified by e.g. (Humphries and Woodland, 1997; Cremelie and Martens, 1999; Riley et al., 1999; Amdal et al., 2000a,b; Yang et al., 2002; Goronzy et al., 2004; Raux, 2004; C. Van Bael and Strik, 2007). All data-driven approaches have the following two steps in common: (1) the collection of (base form, observed transcription) pairs that reveal the relevant transformation mechanisms one would like to model for the application domain (e.g. spontaneous speech recognition), (2) the automatic derivation of rules that can transform a base form to a set of transcriptions which, on average, comprises a transcription that is better approaching the observed transcription than the base form.

Ideally, the observed transcriptions of spoken utterances are obtained manually by listening to these utterances and by writing down the heard phonetization, called the auditorily verified (AV) transcription. Unfortunately, gathering AV transcriptions for a large and representative set of training utterances is very costly.

This explains why only a few studies (e.g. (Riley et al., 1999)) actually use such transcriptions.

A more practical approach is to use the acoustic models to retrieve the most likely pronunciation and to consider this pronunciation as the observed one. In that case one can use a phoneme recognizer incorporating n-gram phonotactics, like in ((Humphries and Woodland, 1997; Amdal et al., 2000a,b; Goronzy et al., 2004)), or alternatively, an aligner that lines up the utterance with an automaton representing the baseline transcription and some deviations (phoneme deletions, insertions, substitutions) thereof, like in (Cremelie and Martens, 1999; Riley et al., 1999; Yang et al., 2002; Raux, 2004). Using the first approach, the observed transcription can deviate considerably more from the base form, which may be appealing in the case of accented non-native speech (because large deviations can be expected). On the other hand, this approach may be less reliable than the second one, which raises the need for good confidence measures to exclude low-quality outputs.

Since the aim of the above strategy is to transform the base form towards the observed transcription, we prefer, from now on, to use the terms source transcription and target transcription instead.

Considering all the experimental results for LVCSR reported in the cited papers, the conclusion seems to be twofold: although lexical modeling is not effective in combination with state-of-the-art cross-word triphone acoustic models for native speech recognition, it can offer a moderate gain (9 to 15% relative) for the recognition of non-native speech.

2.2.2. Lexical modeling for proper name recognition

A very relevant problem to tackle in the context of proper name recognition involving foreign names and/or non-native speakers is the modeling of cross-lingual effects. One popular solution is to include additional G2P transcriptions emerging from G2P converters that cover some of the foreign languages. This way one can exploit the foreign phonological knowledge encoded in these G2P converters. In (Cremelie and ten Bosch, 2001) Dutch, English and French G2P transcriptions were included for all entries in a pronunciation dictionary containing about 500 names: Dutch, English, French and other names. Using optimized language dependent weights for these transcriptions, the name error rate could be reduced by about 40% for native Dutch speakers, 70% for French speakers, 45% for English speakers and 10% for other foreign speakers.

A similar approach was adopted in (Maison et al., 2003), but now in a larger scale set-up with a 44K person name vocabulary. Two baseline pronunciation dictionaries were constructed: one with handcrafted typical transcriptions for native speakers and one with transcriptions generated by a native G2P converter. Then, new variants were generated by eight foreign G2P converters covering all the foreign languages occurring in the data set. Using n-gram grapheme models as language identifiers, likelihoods for the source languages of the names were computed and the transcriptions generated by the top 2 foreign G2P converters were added to each of the baseline lexicons. The native language was US English, and the test utterances consisted of (1) US proper names uttered by US native speakers, (2) foreign proper names (Mandarin Chinese, Czech, French, German, Hindi, Italian, Russian or Spanish) uttered by US native speakers and (3) foreign proper names uttered by foreign speakers whose mother tongue belongs to the just mentioned group of foreign languages. The variants caused a 25% reduction of the name error rate for the non-native utterances, irrespective of the baseline. However, the reduction was only 10% for the native utterances of foreign names and insignificant for the native utterances of native names.

In (Li et al., 2007), another relevant problem is tackled. They propose a method to adapt the faulty G2P conversion process for names. The parameters of a so-called graphoneme n-gram transcription model are modified on the basis of acoustic likelihoods calculated by means of recognition experiments on a limited amount of adaptation data. Graphonemes are defined as (graphemic pattern, phonemic pattern) pairs and a graphoneme transcription model is trained using a lexicon of aligned graphemic and phonemic sequences (see e.g. (Bisani and Ney, 2008) for more details). Given a set of name utterances, the name orthography and the accompanying acoustic evidence of each utterance are employed to look for the phonemic sequence that maximizes the sum of the likelihoods emerging from the G2P process and the acoustic models respectively. The found phonemic sequences are then used in combination with the orthographies to update the graphoneme n-gram model. The authors propose a discriminative training algorithm to optimize the conversion process such that it minimizes the name error rate for the training set. With their method they achieved a 12% reduction of the error rate on an independent test set.

The above results clearly demonstrate that lexical modeling can substantially improve the recognition of proper names. The objective of the present work is to further explore this and to assess the factors affecting the gains that can be achieved. Note that this work can be considered as a unification and extension of previous work that we published in a number of conference pa-

pers (van den Heuvel et al., 2009; Réveil et al., 2009, 2010). Note too that we consider the inclusion of multiple G2P transcriptions covering the foreign languages of interest as an established approach, and that we will include this recipe in the baseline system against which we will experimentally validate our own approaches.

3. The proposed lexical modeling methodology

The proposed method creates pronunciation variants on the basis of automatically derived stochastic transformation rules. Each rule predicts with which probability a phoneme sequence (called the *focus*) appearing in the source transcription may be pronounced as an alternative phoneme sequence (called the *rule output*) when it occurs in a particular linguistic context that can be defined in a flexible way (see below). The rules for a certain focus are embedded in the leaf nodes of a binary decision tree which uses yes/no-questions to distinguish between different contexts. Since the rules are stochastic in nature they will lead to multiple solutions per name with different probabilities attached to these solutions.

The present approach constitutes a unique combination of the following features: (1) the transformable objects can be phonemic sequences (phoneme patterns) of different lengths, (2) the linguistic context is not restricted to the phonemic context (as in many other studies) but it can also include orthographic (graphemic), syllabic, morphological, syntactic and semantic information in a flexible way, (3) some general procedures are available to support the computer-aided identification of beneficial syllabic and morphological features, (4) the relevant (focus, output) combinations as well as the rules are learned automatically. Many published methods share some of the mentioned features, but we believe to be the first to propose and to assess a method incorporating all of these features simultaneously.

Another distinctive property of our method is that it only needs a small lexical database of the order of thousand names representative of the envisaged application domain as domain knowledge. Per name, this database has to supply one or more plausible pronunciations and, optionally, some tags further characterizing the name (e.g. the name category, see later). We argue that in many practical situations, such a database can be created cheaply because of its limited size, and because it can be done by one or two persons acquainted with the domain (and able to write phonetics): they can select the names and enter some typical pronunciations of each name, either starting from scratch or from the native G2P transcription of the name.

Since the target transcriptions have to be supplied by a human, the method as a whole is semi-automatic, but once the targets are available, the rest of the method is conceptually fully automatic. Nevertheless, it is practically implemented as a process that permits the user to intervene in an easy and transparent way if he has reasons to believe that he can improve the automatic procedure. As will be explained further, these interventions boil down to simple updates of text files on the basis of statistical information that is being generated automatically after each step of the automatic procedure.

Let us now discuss in detail the different steps of our method, starting with a review of the contextual features we have chosen to include. The software implementation is publically available through the Dutch HLT (Human Language Technology) center as the Autonomata transcription toolkit[2]. This release also contains extensive documentation on the toolkit.

3.1. Contextual features

First of all, we adopt the phonemes immediately preceding and succeeding the focus as the primary contextual features. However, as in (Riley et al., 1999), we also take lexical stress symbols and syllabic information into account, be it in a somewhat different way. We consider the identities of the two phonemes to the left and to the right of the focus (4 features), the identities of the vowels of the focus syllable and its two neighboring syllables (3 features) and the stress levels (no stress, primary stress or secondary stress) of these syllables (3 features).

Secondly, we follow the argument of (Schaden, 2003a,b) that the orthography plays a crucial role in pronunciation variation modeling, in particular in non-native pronunciation variation modeling. Take the French cheese name "Camembert" for instance. While the native pronunciation of this name is /"ka.ma˜.bER/, a native Dutch speaker may be inclined to pronounce it as /ka.m@m."bErt/ because in Dutch, a "t" in the orthography is normally not deleted in the pronunciation (see (Schaden, 2003a) for more examples). The main limitation of Schaden's work was that the rules were handcrafted rules. In a similar vein, (Bouselmi et al., 2006) incorporated graphemic information in an automatic data-driven approach, but the limitation of that work was that the focus had to be a single phoneme and that the graphemic context was restricted to the grapheme that gave rise to this focus. In our terminology, one could also say that the focus has become

[2] See www.inl.nl/en/tools

a (grapheme, phoneme) pair, also called a graphoneme, and that no further graphemic context is allowed. For our experiments, we considered 4 graphemic features: the graphemic pattern that caused the focus (restricted to the first two graphemic units though), the graphemic units immediately left and right of this pattern, and a flag signaling whether or not the graphemic pattern causing the focus ends on a dot (= a simple indicator of an abbreviation). In a recently published paper (Loots and Niesler, 2011) reference was made to our work and evidence was provided that graphemes as contextual features lead to significant gains in the generation of accented South African English transcriptions from American English and British English pronunciations.

Thirdly, we support the intention formulated in (Schaden, 2003a) to employ morphological information as a potentially interesting context descriptor. Schaden noticed for instance that the vowels in the German suffixes "-stein" and "-bach" are less susceptible to accented pronunciations than the same vowels in other morphological contexts, but he did not actually build a system incorporating this aspect. Since a true morphological analysis may be difficult to incorporate, especially in the domain of proper names, we include 7 'morphological' features revealing the syllabic context the focus appears in, and taking the presence of typical name heads and tails (prefixes and suffixes) into account:

1. 3 booleans indicating whether the focus syllable, the previous and the next syllable belong to a user-specified syllable list,
2. a boolean indicating whether the focus appears in a word starting with a prefix that belongs to a user-defined prefix list,
3. a boolean indicating whether the focus appears in a word ending in a suffix that belongs to a user-defined suffix list,
4. the positions (in number of syllables) of the focus start and end w.r.t. the first and last syllable of the name stem respectively (the name stem is obtained by depriving the name of the longest prefix and suffix from the user-defined prefix and suffix lists)[3].

Further below we will explain how to get the mentioned syllable, prefix and suffix lists in an automatic way.

Finally, we believe that in the envisaged applications of proper name recognition, high-level semantic information such as the name category (e.g. street name, city name, Point-of-Interest), the mother tongue of the speaker (if known), the language of origin of the inquired name (if known), etc. are important to create more dedicated pronunciation variants. Therefore, we devised our software so that such semantic tags can be accommodated through boolean features which are true if the tag belongs to predefined value sets (the values are character strings). In the experiments that will be discussed later, we employed the name category as a semantic feature, while the mother tongue of the speaker and the language of origin of the name were used as a means to distinguish between different P2P converters, designed for specific combinations of these variables.

3.2. The overall rule induction process

Since we aim to work with variable length phoneme patterns in the focus and with different types of contextual features in the rule condition, the rule induction process is a little more complicated than usual. The whole process is depicted in Figure 1. In general terms,

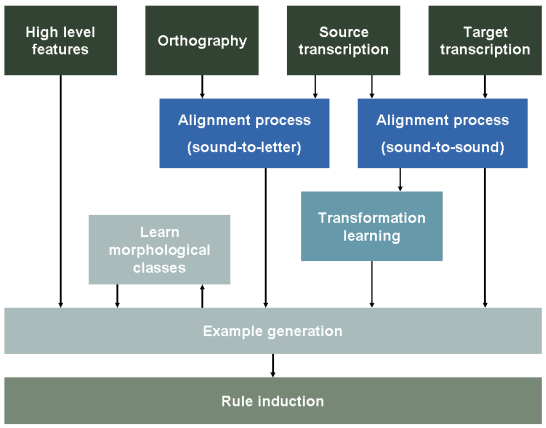

Figure 1: Process for automatically learning of a P2P converter.

the process is applied to a set of training objects each consisting of an orthography, a source transcription (the base form), a target transcription and a set of high-level features. Given these training objects, the learning process then proceeds as follows:

1. The objects are supplied to an alignment process incorporating two components: one for lining up the source transcription with the target transcription (sound-to-sound) and one for lining up the source transcription with the orthography (sound-to-letter). These alignments, together with the (morphological and/or semantic) high-level features are stored in an alignment file.

[3]If the focus starts/ends in the selected prefix/suffix, the corresponding position is zero.

2. The transformation learner analyzes the alignments and identifies the (focus, output) pairs that are capable of explaining a lot of systematic deviations between the source and the target transcriptions. These pairs define transformations which are stored in a transformation file.
3. The alignment file and the transformation file are supplied to the example generator that locates focus patterns from the transformation file in the source transcriptions, and that generates a file containing the focus, the corresponding contextual features and the output for each detected focus pattern. These combinations will serve as the examples from which to train the rules. If no morphological features have been defined yet, one can define them on the basis of information produced by the example generator (see below) and run the example generator a second time to create training examples that also incorporate these features.
4. The example file is finally supplied to the actual rule induction process which automatically constructs a decision tree per focus.

In the subsequent subsections we further elaborate the rule learning process and we also indicate where a manual intervention is possible or desirable.

3.3. The alignment process

As indicated before, the alignment process performs a sound-to-letter (or phoneme-to-grapheme) alignment between the source transcription and the orthography, and a sound-to-sound (or phoneme-to-phoneme) alignment between the source and the target phonemic transcription. By replacing every space in the orthography by the symbol "()", one can visualize both alignments together in the form of a matrix (see Figure 2). The rows subsequently represent the orthography (row 1), the source transcription (row 2) and the target transcription (row 3).

D	i	r	k	()	V	a	n	()	D	e	n	()	B	o		ssch	e		
"	d	I	r	k	#	f	A	n	#	d	E	n	#	"	b	O	.	s	@	
"	d	i	r	k	#	v	A	n	#	d	@	m	#		b	O	.	s	@	

Figure 2: Alignment of the orthography (top), the source transcription (mid) and the target transcription (bottom) of the person name *Dirk Van Den Bossche*.

The alignment between a source transcription and a destination transcription (either the orthography or the target phonemic transcription) is obtained by means of Dynamic Programming (DP). To control the DP process, we first of all define for each source unit u_s, defined as a unit that can appear in the source transcription, a so-called image set $\mathcal{I}(u_s)$. This set comprises all the destination units different from u_s that are likely to be lined up with u_s. Then we define five probabilities:

1. P_d: the chance that a unit of the source transcription is not lined up with any unit of the destination transcription (= deletion),
2. P_i: the chance that a unit of the destination transcription is not lined up with any unit of the source transcription (= insertion),
3. P_{si}: the chance that, in the absence of a deletion/insertion, a unit of the source transcription is lined up with a unit of the destination transcription that belongs to the image set of that source unit (= expected substitution),
4. P_{so}: the chance that, in the absence of a deletion/insertion, a unit of the source transcription is lined up with a unit of the destination transcription that does not belong to the image set of that source unit (= unexpected substitution).
5. P_{eq}: the chance that, in the absence of a deletion/insertion, a unit of the source transcription is lined up with a unit of the destination transcription that is equal to that source unit (= no substitution).

Note that in the case of a sound-to-letter alignment the source phonemic unit can never be equal to the destination graphemic unit, and consequently $P_{eq} = 0$.

If T_{sn} and T_{dm} represent the n-th and m-th element of the source and destination transcription respectively, and if the alignment search space is represented by a trellis, the log probability $L_{n,m}$ for reaching node (n, m) of the trellis is obtained by

$$\begin{aligned} L_{n,m} &= \max[L_{n-1,m} + \log P_d, \\ &\qquad L_{n,m-1} + \log P_i, \\ &\qquad L_{n-1,m-1} + \log(1 - P_d - P_i) + S_{n,m}] \\ S_{n,m} &= \log P_{eq} \quad \text{if } T_{dm} = T_{sn} \\ &= \log P_{si} \quad \text{if } T_{dm} \in \mathcal{I}(T_{sn}) \\ &= \log P_{so} \quad \text{if } T_{dm} \neq T_{sn} \text{ and } T_{dm} \notin \mathcal{I}(T_{sn}) \end{aligned}$$

The DP process is further prohibited to line up a boundary source unit (a syllable or word boundary) with a non-boundary destination unit. In the case of a sound-to-sound alignment, it is also prohibited to line up a stress marker with anything else than a stress marker. In the case of a sound-to-letter alignment, a stress marker cannot be lined up with any graphemic unit.

From the above it follows that for both the sound-to-sound and the sound-to-letter alignment, the user has to

provide the five probabilities P_d, P_i, P_{si}, P_{so} and P_{eq} and the image sets of the units that can appear in the source transcription.

Since it is generally known that certain graphemic patterns (e.g. “eau”, “ie”, “ij”, etc. in Dutch) often give rise to one sound, the sound-to-letter alignment also accommodates that a single source unit is lined up with a sequence of up to 4 graphemes, that is then included as an additional target unit in the image set of the particular source unit (e.g. “eau” in the image set of /o/). Figure 2 shows a multi-character pattern “ssch” which is lined up with the source phoneme /s/.

For the sound-to-sound alignment one can start with minimal image sets, only comprising units differing from the source unit in one of the following properties: long/short (for vowels), voiced/unvoiced, nasalized/not-nazalized, diftong/monophtong). The required probabilities can be initialized to $P_d = P_i = 0.10$, $P_{so} = 0.05$, $P_{si} = 0.15$ and $P_{eq} = 0.80$. For the sound-to-letter alignment one can start with minimal image sets representing the most important context-independent grapheme-to-phoneme rules one could formulate for the language. The probabilities can be initialized to $P_d = 0.05$, $P_i = 0.15$, $P_{so} = 0.15$, $P_{si} = 0.85$ and $P_{eq} = 0$, with $P_i > P_d$ expressing that a phoneme is more often the consequence of multiple graphemes than vice versa.

Since the image sets mostly represent domain independent knowledge, good baseline alignment control files for a certain language can be constructed once, and later be reused for different domains. The user then has the opportunity to update the files manually on the basis of statistical information (most frequently observed sound-to-sound and sound-to-letter substitutions, number of deletions, insertions and substitutions within and outside the image sets) and to repeat the alignments with these new files.

3.4. Transformation retrieval

In a second stage, the outputs of the aligner are analyzed in order to identify the (focus,output) transformations that can explain a large part of the observed discrepancies between the source transcriptions and the corresponding target transcriptions.

Since stress markers are always lined up with stress markers (see previous section), and since every syllable is presumed to have a stress level of 0 (no stress), 1 (secondary stress) or 2 (primary stress), the stress transformations are restricted to stress substitutions. All of the six possible substitutions that occur frequently enough are retained as candidate stress transformations.

The candidate phonemic transformations are retrieved from the computed alignments after removal of the stress markers. That retrieval process is governed by the following principles:

1. consecutive source phonemes that differ from their corresponding target phonemes are kept together to form a single focus,
2. this agglomeration process is not interrupted by the appearance of a matching boundary pair (as we also want to model cross-syllable phenomena),
3. a focus may comprise a boundary symbol, but it cannot start/end with such a symbol (as we only attempt to learn boundary displacement rules, no boundary deletion or insertion rules),
4. (focus,output) pairs are not retained if the lengths of focus and output are too unbalanced (a ratio > 3), or if they imply the deletion/insertion of three or more consecutive phonemes,
5. (focus,output) pairs not passing the unbalance test are split into two shorter candidate transformations whenever possible.

Once all utterances are processed, the set of discovered transformations is pruned on the basis of the phoneme discrepancy counts associated with these transformations. The phoneme discrepancy count expresses how many source phonemes would become equal to their corresponding target phoneme if the transformation were applied at the places where it helps (and not at any other place). In our experiments we always retained all transformations with a phoneme discrepancy count that is larger than 0.5% of the total number of phoneme errors encountered in all training alignments. A phoneme error is said to occur whenever a source and target phoneme are different (deletion, insertion, substitution).

Figure 3 shows one stress transformation (from primary to no stress) and three phonemic transformations (/I/,/i/), (/f/,/v/) and (/E n/,/@ m/) that comply with the five mentioned principles and that emerge from the alignment of Figure 2.

	D	i	r	k	()	V	a	n	()	D	e	n	()		B	o		ssch	e
“	d	I	r	k	#	f	A	n	#	d	E	n	#	“	b	O	.	s	@
“	d	i	r	k	#	v	A	n	#	d	@	m	#		b	O	.	s	@

Figure 3: Candidate transformations that can be retrieved from the alignment of Figure 2.

3.5. Example generation

Now that the relevant transformation list is available, the focuses appearing in that list are used to segment the source transformation of each training object. The

segmentation is performed by means of a stochastic automaton. This automaton represents a unigram model that comprises a set of phoneme consuming branches. Each branch corresponds to a single or multi-state focus model containing states to consume the subsequent phonemic symbols of the focus it represents. One additional branch represents a single-state garbage model that can consume any phonemic unit. The probabilities on the transitions towards the focus models are derived from the number of times the different focuses were transformed in the training data (counters included in the transformation list). The probability for entering the garbage model and the probability of the feedback loop from the final to the initial state are set to 0.1 times the minimum of the smallest forward transition probability to any of the focus models. This guarantees that any one-symbol focus will be preferred over the garbage model and that a multi-state focus model will be preferred over a sequence of single state focus models.

Once the segmentation of a source transcription is available, a training example will be generated for each focus segment encountered in that transcription. Each example consists of a focus, the corresponding output and the contextual features. The latter can be retrieved from the information provided by the aligner. The output must either be equal to the focus or to one of the outputs found in the candidate transformation list, otherwise no training example will be created.

As stated before, the user can define 'morphological' features on the basis of user-specified syllable, prefix and suffix sets. To help the user in defining these sets, the example generator automatically creates a list of syllables that frequently contain a discrepancy between the source and the target transcription. It also generates a list of initial and final name parts (only parts composed of one or two syllables) that frequently co-occur with such a source-target discrepancy. Since the frequencies of co-occurrence are also available, one can easily decide which items to retain in the envisaged syllable, prefix and suffix sets. For our own experiments, we selected all items having frequencies of more than 1% of the total number of syllables and names in the training set respectively.

Once the 'morphological' sets are defined, one can re-run the example generator to generate training examples also comprising the desired 'morphological' features. All these training examples are collected in one training database.

3.6. Rule induction

From the training examples, the system finally learns a decision tree for each focus appearing in the transformation list. The stochastic transformation rules are attached to each of the leaf nodes of such a tree. The identity rule (do not perform any transformation) is one of the stochastic rules in each leaf node. The collection of all learned decision trees constitutes the actual P2P converter.

The decision trees are grown incrementally by selecting at any time the best node split one can make on the basis of a list of yes/no-questions concerning the transformation context and a node splitting evaluation criterion. All questions are defined in the form of symbol sets. The phonemic symbol sets represent the phonological classes normally used for controlling the state tying in acoustic modeling. The graphemic symbol sets also refer to phonological classes (e.g. graphemic symbols usually giving rise to a phoneme of such a class) or to orthographic properties (e.g. digits, special symbols like "ä" or "ë" and punctuation marks). There were also questions about the nucleus of the syllable, or about particular syllable, prefix and suffix subsets of the formerly defined full sets. However, the latter questions did not seem to contribute much to the performance and were finally left out for the experiments described later.

The node splitting evaluation criterion is entropy loss. If N_{xk} represents the number of training examples selecting tree node x and obeying rule k of that node, the entropy of node x is computed as

$$H(x) = -\sum_k N_{xk} \log \frac{N_{xk}}{N_x}, \quad \text{with } N_x = \sum_k N_{xk}$$

If a certain leaf node *leaf* is now examined for a certain question which would create the sub-nodes *yes* and *no* the entropy loss per example originally selecting the leaf node is given by

$$L_H = \frac{H(leaf) - H(yes) - H(no)}{N_{leaf}}$$

The rule induction process is then controlled by three thresholds: the minimum L_H to achieve in order to retain a node split (set to 0.01 here), the minimum number of times a tree node is visited before it can be retained (set to 0.01% of the total number of training examples) and the minimum firing probability a rule must have in order to be included in the rule list of a leaf node (set to 0.1 here).

An example of an actual rule that was learned for Dutch names uttered by English natives is the following:

> The phonemic sequence /d @/ in a Dutch G2P pronunciation can, with a probability of 0.33,

be replaced by a /t/ if it is graphemically preceded by a diphtong (e.g. "au"), not followed by an "r" and if the second phoneme to the right is not an obstruent.

Consider now the following last names and their Dutch G2P transcription: "Van Der Heest" (/vAn#d@r#"he:st/), "Stroders" (/"stro:.d@rs/), "Muidenier" (/"m9y.d@.nir/) and "Oudebrand" (/Au.d@.brAnt/). As a consequence of the rule /Aut.brAnt/ will be created as a variant of the last name, which is very plausible. For the other names, the rule condition is not satisfied.

3.7. Pronunciation variant generation

In generation mode, the trained P2P converter parses the G2P transcription from left to right in the same way as described before. For every focus segment it then applies its rules to generate multiple pronunciation variants with attached probabilities. The procedure is the same as the one explained in (Cremelie and Martens, 1999).

4. Experimental conditions

In order to make an experimental assessment of the proposed methodology, we need a speech recognition engine, a spoken name corpus and an evaluation criterion to score different systems.

4.1. Speech recognition engine

The recognition experiments were conducted with the Dutch version of the commercially available state-of-the-art Nuance VoCon 3200 recognizer[4]. The engine was delivered with two acoustic models:

- AC-MONO: the standard Dutch model, trained on speech of native Dutch speakers from the Netherlands and Belgium. The underlying phoneme set consists of 45 phonemes.
- AC-MULTI: a multilingual acoustic model, trained on the same data as AC-MONO, but supplemented with equally large amounts of UK English, French and German speech. The underlying phoneme set consists of 80 phonemes, and models for phonemes appearing in multiple languages have thus seen data from all these languages.

[4]http://www.nuance.com/for-business/by-product/automotive-products-services/vocon3200/index.htm

The models were combined with a simple name loop and a lexicon with or without pronunciation variants. However, no prior probabilities can be attached to these variants, meaning that the learned rule probabilities cannot be fully exploited.

4.2. Spoken name corpus

All experiments were conducted on the Autonomata Spoken Name Corpus (ASNC). It consists of recordings of Dutch, English, French, Moroccan and Turkish names (person names (first name + surname), street names and city names) spoken by Dutch, English, French, Moroccan and Turkish speakers. Each of the 240 recorded speakers provided 181 utterances (tokens), with every utterance containing one spoken name. In the whole corpus 3540 unique names can be discerned. More details on the corpus composition can be found in (van den Heuvel et al., 2008).

Since cross-lingual effects are anticipated to be important, we have divided the ASNC into smaller subcorpora (called cells) on the basis of the *speaker tongue*, defined as the mother tongue of the speaker, and the *name source*, defined as the language of origin of the name. The cell (DU,EN) for instance contains the recordings of Dutch speakers reading English names. We furthermore made a distinction between the languages that occur in our corpus: we discern Dutch as the native language (as it is the most prominent language in the corpus, both in terms of names as in terms of speakers), English and French as two non-native languages of the so-called NN1 type, and Turkish and Moroccan as two non-native languages of type NN2. The division of the non-native languages can be motivated in two ways. First of all, all native Dutch speakers that read English/French names were familiar with the English/French language, but not necessarily with the Turkish/Moroccan language. Secondly, we had access to English and French G2P converters, but not to Turkish or Moroccan G2P converters.

As we were only interested in situations where not both the speaker tongue and the name source are non-native, and as there was not much sense in keeping Turkish and Moroccan apart, we only considered seven cells in our experiments. Table 1 shows the number of training and test utterances in each of these cells (note that the cell (DU,DU) is listed twice). Care was taken to ensure that there is no overlap in speakers and spoken names between the training and the test set (the two sets are defined in the documentation included in the ASNC distribution).

For investigating the effects of vocabulary size, we have constructed two vocabularies. One vocabulary

Table 1: Number of tokens per (speaker tongue,name source) combination in the ASNC training and test set. The notation (DU,*) refers to the four combinations involving native Dutch speakers. The notation (*,DU) refers to the four combinations involving native names.

	Set	DU	EN	FR	NN2
(DU,*)	train	9960	1909	966	2188
	test	4440	851	414	992
(*,DU)	train	9960	3000	1680	4920
	test	4440	1800	720	2280

consists of the 3540 unique names occurring in the ASNC. Another vocabulary has 21240 entries and is obtained by supplementing the first vocabulary with 17700 extra person names and geographical names. The balance in terms of name source and name type is preserved.

Each name in the corpus comes with a typical Dutch transcription (TY), manually created without listening to any recording. Each name token comes with a Dutch auditorily verified (AV) transcription, made by a human expert after having listened to the utterance as many times as needed.

4.3. Evaluation measures

The recognition accuracy will be expressed as a Name Error Rate (NER), defined as the percentage of incorrectly recognized names. A name is only correct if all its constituents (parts) are correct. Statistical significance of NER differences is determined using the Wilcoxon signed ranks test (Conover, 1999).

During P2P development, we also considered the Transcription Error Rate (TER), defined as the percentage of names for which none of the available lexicon transcriptions is equal to the correct (= auditorily verified) transcription. Since syllable boundaries and stress markers are not that important for recognition, they are ignored in the computation of TER. The relative Transcription Improvement Rate (rTIR), which is defined as the percentage of names for which at least one P2P transcription has a lower Levenshstein distance to the correct transcription than the source transcription, then constitutes a measure for a first quick assessment of the effect of the P2P converter (no time consuming recognition experiment required).

4.4. Modes of operation

Since in certain situations it is plausible to presume prior knowledge of the speaker tongue and/or the name source, three relevant modes of operation of the recognizer will be considered:

M1: In this mode, the speaker tongue and the source of the inquired name are a priori known. I.e. the case of a tourist who uses a voice-driven GPS system to find his way in a foreign country where the names (geographical names, POI names) all originate from the language spoken in that country.

M2: In this mode, the speaker tongue is known but names from different sources can be inquired. Think of the same tourist who is now traveling in a multilingual country like Belgium where the names can either be Dutch, English, French, German, or a mixture of those.

M3: In this mode, neither the mother tongue of the actual user nor the source of the inquired name are a priori known. This mode applies for instance to an automatic call routing service of an international company.

If the recognizer is operated in mode M1, it means that we know in which cell we are. This implies that when variants are added to the lexicon, we will only add them for the entries that can occur in this cell. It is allowed to use cell-specific P2P converters to create the variants then.

5. Experimental results and discussion

The source transcriptions in the subsequent experiments emerge from the Dutch, French or English G2P converters that are used in all Nuance recognition and synthesis products for these languages. We first conduct our experiments under mode of operation M1, but later we also show results for the other modes.

5.1. Effectiveness of standard recipes

The first aim of our baseline experiments is to establish under which circumstances two standard recipes for improving proper name recognition are beneficial. The second aim is to define a good baseline system against which to compare the newly proposed methods. The two investigated standard recipes are: (1) decoding the utterances with a multilingual acoustic model and (2) including foreign G2P transcriptions in the lexicon. Note that when used in combination with a monolingual acoustic model, the foreign G2P transcriptions must be nativized: non-Dutch phonemes must be mapped to Dutch phonemes. When used in combination with a multilingual model, nativizing the G2P transcriptions is not necessary but an option.

Figure 4 summarizes the results of all our baseline experiments. It shows the effects of changing the acoustic

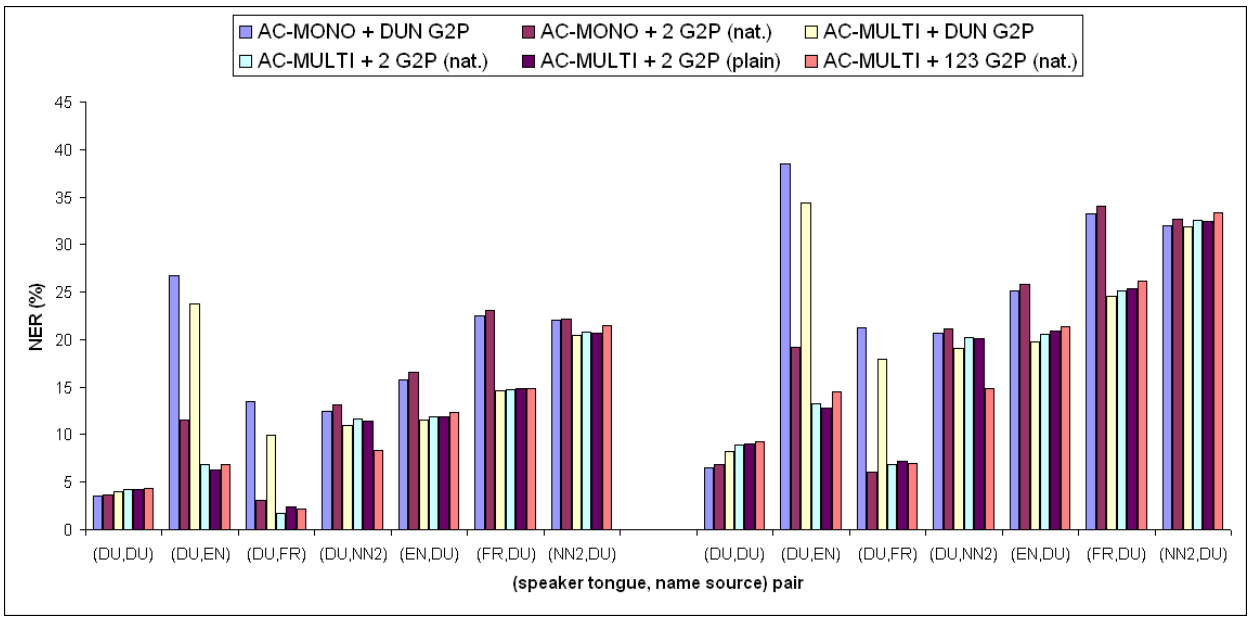

Figure 4: Baseline NER results per ASNC cell for two sizes of the vocabulary: 3.5K (left) and 21K (right). The tested systems are characterized by (a) their acoustic model (monolingual = AC-MONO, multilingual = AC-MULTI), (b) the G2P transcriptions that were included in the lexicon (DUN G2P = only a Dutch G2P transcription per name, 2 G2P = additional English/French transcription for English/French names, 123 G2P = English and French transcriptions also added for NN2 names) and (c) the use of plain or nativized foreign G2P transcriptions.

model, including or excluding foreign G2P converters and using nativization or not.

The data first of all show that for improving the recognition of French and English names spoken by Dutch speakers, adding a French G2P transcription for French names and an English G2P transcription for English names to a lexicon that only contains an initial Dutch G2P transcription for all names (= so-called '2 G2P' lexicon) is much more effective than replacing a monolingual by a multilingual acoustic model. Furthermore, it is also less harming the recognition of Dutch names uttered by Dutch native speakers. The downside of the lexical modeling step is that it slightly harms the recognition of Dutch and NN2 names (as could be expected, since only English and French names receive an additional transcription) whereas a multilingual acoustic model does not. The above comparison of lexical and acoustic modeling effects has not been reported in such detail before. The improvements brought by the foreign G2P transcriptions for the recognition of native speakers reading non-native names of type NN1 confirm that the Dutch native speakers employ their foreign language knowledge when they are confronted with NN1 names. The recognition results also correlate with the fact that the English/French G2P transcription is often closer (Levenshtein metric) to the true (auditorily verified) pronunciation of an English/French name than the Dutch G2P transcription (English G2P transcription is better in 40% of the cases, French G2P transcription even in 60% of the cases).

A second observation is that introducing multilingual acoustic models is clearly a good strategy for improving the recognition of Dutch names spoken by English and French non-natives. Moreover, it never harms the recognition in a cross-lingual setting. Only the recognition of Dutch native speakers reading Dutch names is degraded, as could have been anticipated from (Bartkova and Jouvet, 2006, 2007). Our results definitely confirm that foreign speakers produce accented pronunciations of the Dutch sounds. This observation relates to the so-called *knowledge transfer*, a concept coming from the second language learning domain (You et al., 2005) that states that speakers use their native language knowledge when they come across words from a foreign origin. The fact that the recognition for NN2 speakers is not significantly improved is due to the fact that the multilingual acoustic model did not see NN2 accented sounds during training.

A third observation is that the effects of lexical and acoustic modeling add up very well, the positive as well as the negative effects. This unequivocally proves the complementarity of the two techniques.

A fourth observation that is also new is that nativizing the foreign transcriptions in a cross-lingual setting

has only a negligible effect on the recognition performance obtained with a multilingual acoustic model. We do know however that the use of foreign phoneme symbols is beneficial for the recognition of French/English names uttered by French/English speakers (Réveil et al., 2009).

Finally, we observe that adding French and English G2P transcriptions for the NN2 names (this leads to our '123 G2P' lexicon) helps to improve the recognition of these names without much harming the recognition of the other names. Apparently, Dutch speakers transfer their foreign language knowledge to read Moroccan and Turkish names. This observation is further confirmed by our earlier finding (Réveil et al., 2009) that NN2 names read by Dutch speakers are better recognized if an additional transcription variant created by a German G2P is included for these names (German is also a language many Dutch speakers are familiar with).

On the basis of the above observations, and keeping in mind that the ASNC comes with nativized transcriptions only, we select 'AC-MULTI + 2 G2P (nat)' as a baseline system for assessing the benefits of our lexical modeling approach. For the sake of completeness, we mention that a set-up including all 3 G2P transcriptions for each name in the lexicon (as in (Cremelie and ten Bosch, 2001)), performed worse than the chosen baseline.

5.2. Introducing P2P transcription variants

In this section we investigate under which circumstances the proposed lexical methodology can further enhance the name recognition performance. We first consider the potential of our lexical modeling approach under the ideal circumstance that auditorily verified transcriptions of many domain name recordings (the ASNC training utterances in our case) are available for P2P training. Then we move to the targeted situation where only typical transcriptions of around thousand names from the application domain are available.

5.2.1. How many variants per entry?

The aim of this first experiment is to identify an upper bound on the number of P2P transcription variants that should be added per vocabulary entry. To that end we train one P2P converter per investigated cell. The training is performed on the (Dutch G2P, AV) transcription pairs retrieved from the ASNC training utterances in that cell.

The P2P converters were first evaluated at the transcription level. In Figure 5, the rTIR is plotted as a function of the maximum number of retained P2P transcriptions per entry for the training and the test set utterances. Transcriptions are ranked according to their probability (a consequence of the rule probabilities) and only the most likely ones are retained.

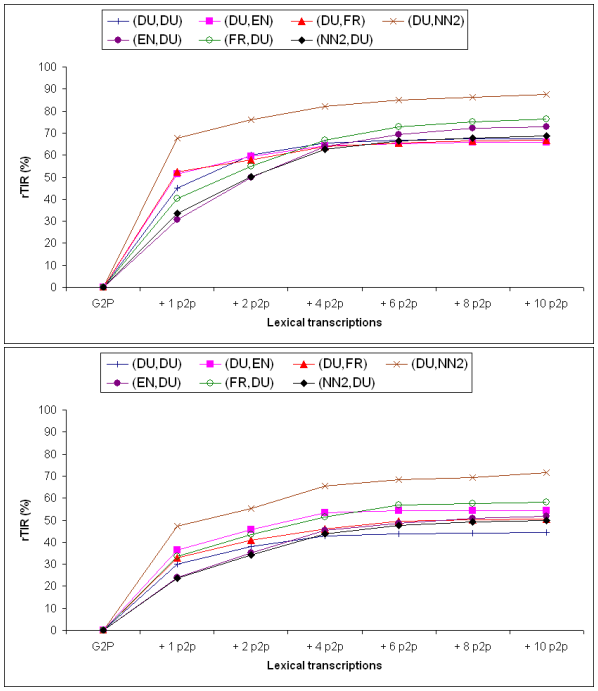

Figure 5: rTIR as a function of the maximum number of P2P variants per entry on the training (above) and test (below) parts of the different ASNC database cells.

The data show that under the given circumstances, P2P converters can generate good variants but that there is no need to add more than four P2P variants per entry to the lexicon. For the training utterances, the best P2P variant has a smaller Levenshtein distance to the auditorily verified target transcription than any of the transcriptions of the 2 G2P baseline lexicon in 65 to 90% of the cases (depending on the investigated cell). For test utterances this percentage lies between 40 to 70%.

As expected, the gains are pretty high for the training set, but recalling that there is no overlap between the test set and the training set names, the P2P rules do seem to generalize well to new names.

5.2.2. Effectiveness of the variants

As shown in Figure 6, utilizing seven P2P converters to add a maximum of four variants per entry to the baseline lexicon leads to significant NER reductions. The most substantial improvement (40% relative) is obtained for the case of Dutch speakers reading NN2 names, but except for cells (DU,DU), (DU,EN) and (DU,FR) in a small vocabulary set-up, all other im-

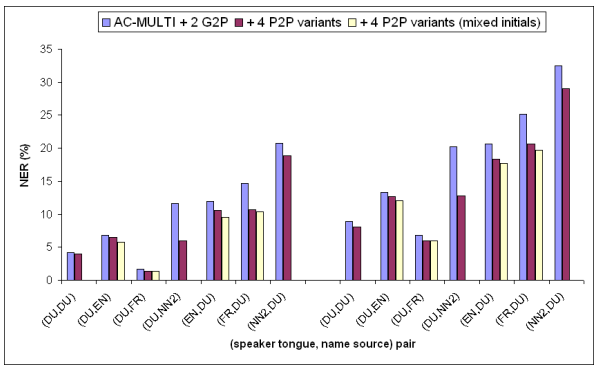

Figure 6: NER results per (speaker tongue, name source) cell for two sizes of the vocabulary: 3.5K (left) and 21K (right). The tested systems all embed a multilingual acoustic model but different lexicons. The baseline lexicon (2 G2P) includes a Dutch G2P transcription for all names and an extra English/French transcription for the English/French names. The lexicon of the second system additionally includes a maximum of four P2P variants. The P2P transcriptions in a cell were produced by a cell-specific P2P converter that was trained on auditorily verified transcriptions. The third system (only relevant for cells involving French or English names or speakers) uses two P2P converters to create the four variants: one that departs from the Dutch and one that departs from the English resp. French G2P transcription.

provements are statistically significant at the 95% level as well.

Apparently, the P2P converter can discover reading strategies that systematically deviate from the one embedded in the Dutch G2P transcriber to produce valuable pronunciation variants. An important rule that was discovered for the cell (FR,DU) for instance, is the following:

> A phoneme /x/ can, with a probability of 0.25, be replaced by a /k/ if the preceding phoneme is not an /r/ or an /s/, if the phoneme 2 positions to the left is not a fortis consonant (/t/, /p/, /k/, /s/, /f/, /S/), if the second phoneme to the right is not a central consonant and if the previous syllable does not carry the primary stress.

In spite of its low firing probability the rule was responsible for several useful variants for street names ending in "-weg". While the baseline Dutch G2P transcription for "Velmseweg" is /"vElm.s@.wEx/, two out of three native French speakers pronounced something that is closer to /"vElm.s@.wEk/.

Note that the NERs are substantially lower than the TERs (not reported). For the Dutch names uttered by Dutch native speakers for instance, the TER after having added up to four P2P variants was still as high as 35%, whereas the NER is less than 5% (similar differences also hold for the other cells). This clearly proves that not having the actually used transcription in the lexicon does not necessarily lead to an incorrect recognition of a name utterance. Conversely, not all transcription improvements reflected in the rTIRs also increase the recognition accuracy.

5.2.3. Foreign language knowledge and knowledge transfer

As argued before, someone's pronunciation of a foreign name can strongly depend on his proficiency of the foreign language in question. For that reason the P2P training set-up in which the source transcription always comes from the Dutch G2P converter is possibly suboptimal. If a native Dutch speaker uses his English language knowledge for instance, his pronunciation of an English name may be closer to that emerging from the English G2P converter. Similarly, if an English speaker uses his mother tongue knowledge to utter a Dutch name (knowledge transfer), his pronunciation may have similarities with the transcription provided by an English G2P.

Hence, for the cells referring to an NN1 name or speaker, there are arguments for training two P2P converters: one that starts from the Dutch G2P converter and one that starts from the G2P converter of the considered NN1 language. We have evaluated two scenarios (per cell): (1) just replace the old P2P converter by the new one, and (2) pool the variants produced by the old and the new converter. In the latter case, each P2P converter is allowed to generate four transcriptions from which the four most probable ones are retained.

The first scenario did not help but the second one did lead to small yet consistent additional gains (see third bar in Figure 6). We therefore retained this scenario in the forthcoming experiments.

5.2.4. Using typical transcriptions as targets

Now that we know what the methodology can achieve under favourable circumstances, it is time to assess its power under the constraint that only typical transcriptions of around thousand proper names are available as target transcriptions for P2P converter training. To that end we employed the typical native Dutch transcriptions from the ASNC training names. Note that since we only possess one typical transcription per name, the P2P converters for the four (*,DU) cells will actually be the same now. Consequently, we could not fully exploit the knowledge that is usually available in mode M1 here. The only thing we can exploit for non-native speakers is the fact that we know the requested name will be Dutch.

Obviously, the P2P rules can no longer model phenomena causing variations in the pronunciation of the same name, as there is only one typical target per name. However, they can still learn to account for discrepancies between the G2P transcription and that typical transcription appearing in particular linguistic contexts.

Since the number of unique names per cell is often small (there are 1676 Dutch names, but only 322 English names, 161 French names and 371 NN2 names), we also performed an experiment in which the training sets for the least populated cells (English and French names) were extended with 684 additional English and 731 additional French names (person names and geographical names) respectively. It was checked of course that none of the added names belonged to the test set.

Figure 7 shows NER results for the baseline lexicons and for the lexicons comprising a maximum of four additional P2P variants. Since the trends are similar for

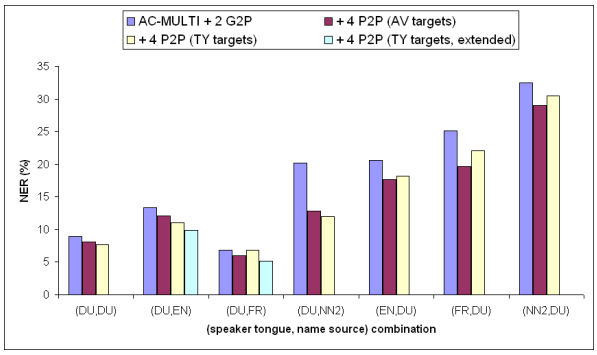

Figure 7: NER results per (speaker tongue, name source) cell for a vocabulary of 21K names. The tested systems all embed the same multilingual acoustic model but different lexicons. The lexicons of the second and the third system are obtained by supplementing the baseline lexicon (2 G2P) with a maximum of four P2P variants. These variants are produced by cell-specific P2P converters that were trained on AV transcriptions (system 2) and TY transcriptions (system 3). The fourth bar shows the results we obtained for the (DU,FR) and (DU,EN) cells with P2P converters that were trained on the typical transcriptions found in the extended training material.

the two vocabulary sizes, only the results for the large vocabulary are reported here. Furthermore, the subsequent discussion refers to systems trained on the extended training set.

A bit surprisingly, the typical transcriptions yield larger NER reductions than the auditorily verified transcriptions in all cells comprising Dutch speakers. We will investigate the reason for this result in the next section. For the recognition of Dutch names spoken by non-native speakers, the typical transcriptions are inferior. There are two phenomena that may have contributed to this finding. One is that we only had typical pronunciations for Dutch native speakers (as stated before). Another is that there is a lot of variation in the pronunciations of a Dutch name by non-native speakers and that this variation is no longer visible during P2P training. We plan to investigate the relative importance of these two phenomena, because if the former one is important, we could possibly further improve our results by also including typical pronunciations of non-native speakers in our lexical training data.

The main conclusion that can be drawn from Figure 7 is that our methodology, when applied on the extended set of typical name transcriptions, yields a statistically significant (at the 99% level) reduction of the NER for all cells referring to native speakers. The improvements amount to almost 15% for the pure native case and to about 25% relative (or higher) for the case of native speakers uttering foreign names. For the case of non-native speakers reading Dutch names, the gain is now just over 6% for NN2 speakers and close to 12% for NN1 speakers. The latter gain is still statistically significant (at the level of 95%).

At the transcription level, the TERs and rTIRs obtained with the new P2P converters (curves not shown here) are now very similar for training and test set. This was anticipated since the P2P converters are no longer learned on actual training utterance transcriptions. Furthermore, the TERs and rTIRs for the test set are very much in line with those obtained earlier with P2P converters trained towards AV targets.

5.2.5. Analysis of recognition improvements

Our first hypothesis concerning the good results obtained with TY targets for native speakers was that for these speakers, there is not that much variation to model within a cell. Hence, one single TY transcription target per name might be sufficient to learn good P2P converters. To verify this hypothesis we measured, per cell, the percentage of utterances for which the auditorily verified transcription was not included in the baseline G2P lexicon. We found percentages of 33% for cell (DU,DU), around 50% for (DU,EN) and (DU,FR) and around 75% for all other cells. Consequently, the measurements support our hypothesis for all cells except (DU,NN2).

In order to find an explanation for the good results in that last cell, we have recorded how many times the two P2P converters (trained with typical and auditorily verified transcription targets respectively) correct the same recognition error and how many times only one of them does. Figure 8 shows the results for the four cells comprising Dutch speakers. It is remarkable that in cell (DU,NN2) the percentage of errors being corrected by

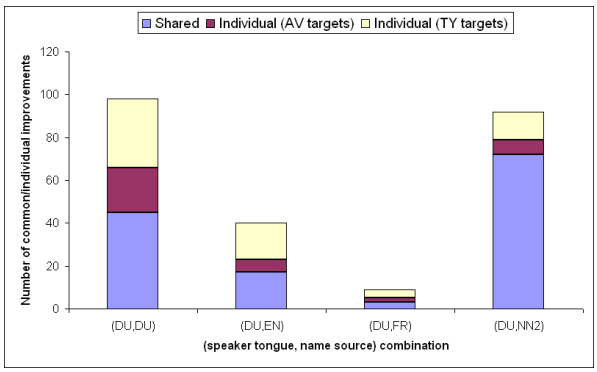

Figure 8: Number of common/individual improvements per (DU,*) due to P2P converters trained on TY and AV targets (21K vocabulary)

both P2P converters is significantly larger than in the other cells. Digging deeper, we came to the conclusion that most of these common corrections were caused by the presence of a small number of simple vowel substitution rules that are picked up by both P2P converters because they represent systematic discrepancies between base forms and typical transcriptions. The most decisive rules express that the frequently occurring letter 'u' in NN2 names (e.g. Curukluk Sokagi, Butrus Benhida, Oglumus Rasuli, etc.) is often pronounced as /u/ (as in "b**oo**t") while it transcribed as /Y/ (like in "m**u**d") or /y/ (like in the French "écr**u**") by the G2P converter.

Motivated by the success of our detailed analysis of what happens in cell (DU,NN2) we have also inspected what happens in the other cells with native speakers after adding P2P variants. Table 2 gives some representative examples of names that were more often correctly recognized with than without P2P variants.

In the table, one cannot observe any real qualitative distinction between the individual errors that either one or both P2P converter types correct. Furthermore, one can see that in the case of shared corrections the best P2P variants generated by each converter are usually very much alike. It appears that the auditorily verified transcriptions do not reveal many pronunciation variation mechanisms that are not already encoded in the typical transcriptions. This is confirmed by an experiment in which the variants generated by the two converters were merged. This merge did not lead to any significant improvement of the recognition accuracy in any of the (DU,*) cells.

An interesting finding (Table 2) is that a minor change in the name transcription (one or two phoneme modifications) can make a huge difference in the recognition accuracy. In this respect, we also have noticed that Dutch speakers have a tendency to over-pronounce certain names. The insertion of an /n/ in the pronunciation of "Duivenstraat" for instance leads to five corrected errors out of six occurrences. Another example of this phenomenon is "Kerkdijk" where the phonologically motivated and correct /"kErg.dEik/ is overruled by the more articulated /"kErk.dEik/.

5.2.6. Effectiveness of variants in mode M2

So far, we assumed that the recognizer has knowledge of the mother tongue of the user and the origin of the name that will be uttered (mode M1). Here, we will examine the effectiveness of P2P variants when names of different origins need to be recognized by a single system that was developed for a specific speaker group (mode M2).

We have only tested mode M2 for the case of Dutch speakers because we do not consider non-native utterances by non-native speakers here, and because we do not possess typical transcriptions of names spoken by non-native speakers anyway. Hence, we performed tests on all cells (DU,*), but with the variants emerging from the P2P converters trained towards TY transcriptions (see previous section) now added simultaneously to the lexicon. The results of these experiments are compared with the corresponding results for M1 and the baseline results obtained with the 2 G2P lexicon in Table 3.

Table 3: NER results (%) for (DU,*) per name source category, baseline results for 2 G2P lexicon compared to results obtained with 4 additional P2P variants from converters trained towards typical training targets, P2P variants are added separately (mode M1) or simultaneously (mode M2) (21K vocabulary)

(DU,*)	**DU**	**EN**	**FR**	**NN2**	**All**
Only G2P	8.9	13.3	6.8	20.2	11.0
+P2P, mode M1	7.7	9.9	5.1	12.0	8.4
+P2P, mode M2	8.4	10.3	5.6	13.2	9.2

For (DU,DU) 60% of the initial gain is lost, but for the other (DU,*) cells the gain is much better preserved. Our first hypothesis was that the larger loss for (DU,DU) was due to the fact that Dutch names have less transcription variants (3.3 on average) than non-native names (4.6 or more on average). However, by allowing more variants for Dutch names we did not get any improvement. On the other hand, by reducing the number of variants for the foreign names, the Dutch names were better recognized again, while the foreign name recognition degraded. Apparently, the necessary Dutch variants are present in the lexicon, but the P2P con-

Table 2: Examples of proper names for which the recognition improves. Listed are: (a) the name, (b) its baseline transcription(s), (c) the P2P variant that led to an error reduction, (d) the P2P converters that that produced it (labeled as TY or AV referring to the targets on which it was trained), (d) the number of times the recognition of a test utterance of the name was improved due to that variant w.r.t. the number of occurrences of the name.

Proper name	**Baseline G2P transcription(s)**	**Helping P2P variant**	**From which P2P?**	**Correction ratio**
Duivenstraat	“d9y.v@.stra:t	“d9y.v@n.stra:t	AV/TY	5/6
Berendrecht	b@.“rEn.drExt	“be:.rEn.drExt	TY	4/6
Kerkdijk	“kErg.dEik	“kErk.dEik	AV	3/6
Carter Lane	“kAr.t@r#“la:.n@ “kA.t@#“le:jn	“kAr.t@r#“le:.n	AV/TY	3/6
Norfolk	nOr.“fOlk “nO.f@k	“nOr.fOk	AV/TY	3/6
Middlesbrough	“mIt.l@z.brux “mI.d@lz.br@	“mI.d@lz.bro:	TY	2/6
Myra Emrick	“mi.ra:#‘Em.rIk “mI.r@#‘Em.rIk	“mAj.“ra:#Em.rIk	AV	2/6
Engreux	EN.“r2:ks a˜.“gr2:	EN.“r2:	AV/TY	2/6
Renée Bastin	r@.“ne:#bAs.“tIn r@.“ne:#ba:s.“te˜	rE.“ne:#bAs.“te˜	TY	3/6

verters that generate variants for English, French and NN2 names induce foreign name variants that are indispensable for a better recognition of the foreign names, but that raise the lexical confusability with the Dutch names.

A plausible alternative approach for the above implementation of mode M2 would be to learn a single P2P converter on all (DU,*) training data, using the name origin as a semantic feature. We actually did a single experiment and we also saw the name origin effectively appearing in the context of some of the learned rules, but the error reduction was smaller than before. We plan to explore this further.

5.2.7. Effectiveness of variants in mode M3

In case neither the mother tongue of the speaker nor the origin of the name is given beforehand (mode M3), the recognition task becomes even more challenging.

Since we have no typical non-native pronunciations of Dutch names at our disposal (these would have been more difficult to obtain because they have to come from non-native speakers with sufficient knowledge of the application domain), a fully realistic evaluation of mode M3 is not possible. However, in Section 5.2.4 we learned that it does not make much difference whether the targets are typical Dutch transcriptions or auditorily verified transcriptions for the training of a P2P converter for foreign names uttered by Dutch natives. Therefore, we argue that using the auditorily verified transcriptions as training targets for the non-native speaker cells is bound to yield a good simulation of what would be possible with non-native typical pronunciations of Dutch names.

Based on the findings of the previous section, we again used our four P2P converters for (DU,*) and we added three more, that were trained with the AV transcriptions found in the (EN,DU), (FR,DU) and (NN2,DU) training cells respectively. Each P2P converter was allowed to generate a maximum of four P2P variants. The variants of the P2P converters developed for the (*,DU) cells were pooled and the four most probable ones were retained in the lexicon. All transcription variants were added simultaneously.

As we employed the auditorily verified transcriptions as targets for the P2P training in the non-native speaker cells now, the M1 results for these cells in Table 4 are as well those obtained with auditorily verified targets. The M1 results depicted for the (DU,*) cells are of course the ones obtained with typical training targets again.

The table demonstrates that the gains (in % relative) attained in mode M1 are pretty well preserved in all cells except for (DU,EN), (DU,FR) and (FR,DU). If we compare the results for modes M2 and M3 for the (DU,*) cells in particular, we notice that the recognition for Dutch names improves again, while the recognition for the non-native names degrades. This is no surprise,

Table 4: NER results (%) per name source category, baseline results for 2 G2P lexicon compared to results obtained with 4 additional P2P variants, in different modes of operation: (M1) variants added separately, (M2) variants added simultaneously per speaker group, (M3) variants added simultaneously for all names and speaker groups. Variants for (DU,*) emerge from P2P converters trained towards typical name transcriptions, variants for non-native utterances of Dutch names from converters trained towards auditorily verified transcriptions (21K vocabulary)

(DU,*)	**DU**	**EN**	**FR**	**NN2**	**All**
Only G2P	8.9	13.3	6.8	20.2	11.0
+P2P, mode M1	7.7	9.9	5.1	12.0	8.4
+P2P, mode M2	8.4	10.3	5.6	13.2	9.2
+P2P, mode M3	7.9	11.0	7.0	13.9	9.1
(*,DU)	**DU**	**EN**	**FR**	**NN2**	**All**
Only G2P	8.9	20.6	25.1	32.5	18.3
+P2P, mode M1	7.7	17.7	19.7	29.0	15.8
+P2P, mode M3	7.9	18.3	22.5	29.6	16.4

as the M3 lexicon is merely an extension of the M2 lexicon, in which additional variants occur for the Dutch names (average number of variants for Dutch names goes from 3.3 to 4.9). These extra variants stem from P2P converters that are trained to model pronunciation variation phenomena for the non-native language speakers. They are thus more likely to maliciously interfere with the variants created for the non-native names. The fact that the gain for (FR,DU) is halved, whilst it is not for the other (*,DU) cells, may be assigned to the fact that the necessary variants for the French utterances of Dutch names are less often in the retained set of 4 P2P variants for Dutch names.

Overall however, one can state that even in modes M2 and M3, a substantial fraction of the gain that was achieved in mode M1 is preserved. This emphasizes the relevance of lexical modeling for large vocabulary proper name recognition in general.

5.2.8. Comparison to other approaches

A direct comparison to other work is difficult due to differences in the recognition engine, the recognition task (isolated proper name recognition in our case) and the corpus on which our method was evaluated. Nevertheless, we try to determine the merits of our work in this section.

One of the unique properties of our approach is that it makes extensive use of non-phonemic features whereas most other approaches do not. Therefore, we made an assessment of the added value of the non-phonemic features in the rule context by training new P2P converters (towards TY targets), but this time without using all the non-phonemic features. We investigated four additional set-ups:

1. only phonemic features used,
2. one graphemic feature added (the pattern that generated the focus),
3. all graphemic features added,
4. only morphological features added

The second set-up gives us an opportunity to apply a version of our method that is closely related to the approach proposed by (Bouselmi et al., 2006) in the context of non-native continuous speech recognition. The major remaining differences are the differences in the transcription alignment strategy and in the definition of the graphemic feature which is constrained to one grapheme in (Bouselmi et al., 2006).

Table 5: NER results (%) for (DU,*) per name source category, 4 additional P2P variants from converters that use different feature sets, mode M1 (21K vocabulary)

P2P features	**DU**	**EN**	**FR**	**NN2**	**All**
None (only G2P)	8.9	13.3	6.8	20.2	11.0
Only phonemic	8.1	10.8	5.8	12.5	9.0
One graphemic	7.9	10.0	4.8	12.1	8.6
All graphemics	8.0	9.7	5.6	11.7	8.6
Morphology	7.8	10.0	4.8	12.0	8.5
All features	7.7	9.9	5.1	12.0	8.4

The results in Table 5 show that the non-phonemic features, as a whole, are beneficial. If we compare the performance of our fully equipped converters to that of converters using only phonemic context features, we establish that in general the latter converters attain only 75% of the gain we achieved with the full converters. On the other hand, one can see that with the inclusion of one graphemic feature it is already possible to reach close-to-optimal performance, at least in our case where that pattern can be composed of multiple graphemes.

The results in Table 5 also show the high potential of the morphological features. They are more helpful than the graphemic features, especially for the economically most important cell (DU,DU).

Another comparison with different techniques can be performed at the transcription level. There, an alternative to the proposed G2P - P2P tandem for creating good proper name transcriptions would have been to collect a large number of names from the application domain and to train a full-fletched G2P converter that can also generate multiple transcriptions for each name.

We trained two such converters for the domain of geographical names. For that training we had a database of 27K examples (pairs of an orthography and a typical transcription) at our disposal. The first system was trained with the Nuance proprietary decision tree learning software, the second system was trained with TiMBL (Daelemans and van den Bosch, 2005), a tool implementing memory-based learning. We performed an evaluation at the phoneme level, and measured the rTIR of the top-1 transcription of each system with respect to the G2P transcription. The rTIR was 75% for the G2P - P2P tandem, 37% for the Nuance decision tree system and 46% for the TiMBL system. Furthermore, the P2P converter was also at least 10 times more compact (memory) than the other two and it was trained on only 2K names. For that number of names, the alternative systems were not much better than the baseline G2P anymore, meaning that these alternatives are not applicable unless a large number of names and their typical transcriptions is available for system training.

6. Conclusions and future work

In this paper we have assessed some existing acoustic and lexical modeling strategies for large vocabulary proper name recognition in a monolingual and cross-lingual setting. We also proposed a new lexical modeling methodology that was evaluated in comparison to these existing strategies in different settings.

Our assessment of existing methodologies demonstrated that in a cross-lingual setting, proper name recognition can benefit a lot from a multilingual acoustic model and from transcriptions emerging from foreign G2P transcribers. Our experiments also showed for the first time that lexical modeling is in some cases more effective than acoustic modeling. We have furthermore witnessed that the two strategies are indeed complementary. Another new finding is that there is no need to use foreign phonemes in a cross-lingual setting with multilingual acoustic models: nativized phonemic transcriptions perform equally well.

The newly presented lexical modeling approach is unique because it combines a number of interesting properties that, for as far as we know, have never been integrated in a single system. Some of these features are: the transformation of variable length phonemic patterns from a baseline transcription, the extensive use of linguistic context at multiple levels (from phonemic to semantic), the computer-assisted identification of syllabic and morphological features, the automatic learning of context-dependent stochastic rules embedded in multiple decision trees, etc. Another feature of the method is that it does not need any labeled speech data as training material nor any expertise in automatic speech recognition. The downside is of course that the user must provide a lexical database of correspondences between a name and its typical transcription. However, since the required database is small (of the order of thousand names), it is easy and cheap to construct.

The new method was evaluated under three modes of operation differing in the a priori knowledge one can assume regarding the mother tongue of the speaker and the language of origin of the name the recognizer will have to process.

When both languages are a priori known, one can achieve important reductions of the name error rate: close to 15% for native proper name recognition, about 25% relative for native speakers uttering foreign names and 12% for foreign speakers uttering native names, provided the mother tongue of the foreign speaker has been involved in the training of the acoustic model.

When names from diverse origins can be inquired, or when multiple speaker groups have to be simultaneously accommodated, a substantial part of the above recognition gains are preserved. One of our future objectives is to study the degradations in detail and to propose solutions that can help us to preserve more. Some preliminary experiments have already been performed, but no breakthrough was achieved so far.

Another objective is to improve the quality of our context feature set: some of the present features seem to be obsolete, whereas we may have overlooked features that are really effective. In that respect we refer to the work of (Schraagen and Bloothooft, 2010), where name utterance repetitions coming from actual users were investigated. It was found that more structural adjustments to previous utterances (e.g. correcting 'sloppy' or incorrect pronunciations) cover 40% of the cases in which a repetition leads to a correct recognition after all. Trying to model these types of structural pronunciation mistakes might be an interesting next phase.

Finally, we argue that the use of priors for transcription variants (not possible within the recognition engine we applied) might further enhance the recognition performance, as we have already established this to be helpful in an experiment that made use of a rescoring of N-best hypotheses.

7. Acknowledgments

The presented work was carried out in the context of two research projects: the Autonomata Too project, granted under the Dutch-Flemish STEVIN program, and the TELEX project, granted by Flanders FWO.

References

Adda-Decker, M., Lamel, L., November 1999. Pronunciation variants across system configuration, language and speaking style. Speech Communication 29 (2-4), 83–98.

Amdal, I., Korkmazskiy, F., Surendran, A., September 2000a. Data-driven pronunciation modelling for non-native speakers using association strength between phones. Proceedings ISCA ITRW ASR2000, 85–90.

Amdal, I., Korkmazskiy, F., Surendran, A., October 2000b. Joint pronunciation modelling of non-native speakers using data-driven methods. Proceedings ICSLP, 622–625.

Bartkova, K., Jouvet, D., May 2006. Using multilingual units for improving modeling of pronunciation variants. Proceedings ICASSP, 1037–1040.

Bartkova, K., Jouvet, D., October 2007. On using units trained on foreign data for improved multiple accent speech recognition. Speech Communication 49 (10-11), 836–846.

Bisani, M., Ney, H., May 2008. Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication 50, 434–451.

Bonaventura, P., Gallocchio, F., Mari, J., Micca, G., 1998. Speech recognition methods for non-native pronunciation variants. Proceedings ESCA Workshop on Modeling Pronunciation Variation for ASR, 17–22.

Bouselmi, G., Fohr, D., Illina, I., Haton, J., May 2006. Fully automated non-native speech recognition using confusion-based acoustic model integration and graphemic constraints. Proceedings ICASSP, 345–348.

C. Van Bael, L. Boves, H. v. d. H., Strik, H., 2007. Automatic phonetic transcription of large speech corpora. Computer, Speech and Language 21, 652–668.

CMU, 2010. Carnegie mellon university pronouning dictionary. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.

Conover, W., 1999. Practical Nonparametric Statistics. Vol. 3. John Wiley & Sons, Inc., New York.

Cremelie, N., Martens, J.-P., November 1999. In search of better pronunciation models for speech recognition. Speech Communication 29 (2-4), 115–136.

Cremelie, N., ten Bosch, L., August 2001. Improving the recognition of foreign names and non-native speech by combining multiple grapheme-to-phoneme converters. Proceedings ISCA ITRW on Adaptation Methods for Speech Recognition, 151–154.

Daelemans, W., van den Bosch, A., 2005. Memory-based language processing. Vol. 1. Cambridge University Press, Cambridge, UK.

Fosler-Lussier, E., Amdal, I., Kuo, H.-K. J., 2005. A framework for predicting speech recognition errors. Speech Communication 46, 153–170.

Gauvain, J.-L., Lee, C.-H., April 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Transactions on Speech and Audio Processing, Vol. 2, 291–298.

Goronzy, S., Rapp, S., Kompe, R., January 2004. Generating non-native pronunciation variants for lexicon adaptation. Speech Communication 42.

Humphries, J., Woodland, P., September 1997. Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition. Proceedings Eurospeech, 317–320.

Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Xiuyang, Y., Sen, Z., 2001. What kind of pronunciation variation is hard for triphones to model. Proceedings ICASSP, 577–580.

Lawson, A., Harris, D., Grieco, J. J., September 2003. Effect of foreign accent on speech recognition in the nato n-4 corpus. Proceedings Eurospeech, 1505–1508.

Leggetter, C., Woodland, P., April 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Computer Speech and Language, 171–185.

Li, X., Gunawardana, A., Acero, A., December 2007. Adapting grapheme-to-phoneme conversion for name recognition. Proceedings ASRU, 130–135.

Loots, L., Niesler, T., January 2011. Automatic conversion between pronunciations of different english accents. Speech Communication 53 (1), 75–84.

Maison, B., Chen, S., Cohen, P., 2003. Pronunciation modeling for names of foreign origin. Proceedings ASRU, 429–434.

Mayfield-Tomokiyo, L., Waibel, A., September 2001. Adaptation methods for non-native speech. Proceedings Workshop on multilinguality in Spoken Language Processing.

PMLA, September 2002. Proceedings itrw on pronunciation modeling and lexicon adaptation for spoken language technology. http://www.isca-speech.org/archive_open/pmla/index.html.

Raux, A., October 2004. Automated lexical adaptation and speaker clustering based on pronunciation habits for non-native speech recognition. Proceedings Interspeech, 613–616.

Réveil, B., Martens, J.-P., D'Hoore, B., September 2009. How speaker tongue and name source language affect the automatic recognition of spoken names. Proceedings Interspeech, 2995–2998.

Réveil, B., Martens, J.-P., van den Heuvel, H., May 2010. Improving proper name recognition by adding automatically learned pronunciation variants to the lexicon. Proceedings LREC, 2149–2154.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G., November 1999. Stochastic pronunciation modelling from hand-labelled phonetic corpora. Speech Communication 29 (2-4), 209–224.

Schaden, S., August 2003a. Generating non-native pronunciation lexicons by phonological rules. Proceedings ICPhS, 2545–2548.

Schaden, S., April 2003b. Rule-based lexical modelling of foreign-accented pronunciation variants. Proceedings 10th EACL Conference, 159–162.

Schraagen, M., Bloothooft, G., May 2010. Evaluating repetitions, or how to improve your multilingual asr system by doing nothing. Proceedings LREC, 612–617.

Stemmer, G., Nöth, E., Niemann, H., 2001. Acoustic modeling of foreign words in a german speech recognition system. Proceedings Eurospeech, 2745–2748.

Strik, H., Cucchiarini, C., November 1999. Modeling pronunciation variation for asr: a survey of the literature. Speech Communication 29 (2-4), 225–246.

Van Compernolle, D., August 2001. Recognizing speech of goats, wolves, sheep and ... non-natives. Speech Communication 35, 71–79.

van den Heuvel, H., Martens, J.-P., D'Hoore, B., D'Haene, C., Konings, N., May 2008. The Autonomata Spoken Name Corpus. Proceedings LREC, 140–143.

van den Heuvel, H., Réveil, B., Martens, J.-P., September 2009. Pronunciation-based asr for names. Proceedings Interspeech, 2991–2994.

Wester, M., Kessens, J., Strik, H., October 2000. Pronunciation variation in asr: which variation to model? Proceedings ICSLP, 488–491.

Yang, Q., Martens, J.-P., Ghesquiere, P.-J., Compernolle, D. V., September 2002. Pronunciation variation modeling for asr: large improvements are possible but small ones are likely to achieve. Proceedings PMLA, 123–128.

You, H., Alwan, A., Kazemzadeh, A., Narayanan, S., September 2005. Pronunciation variations of spanish-accented english spoken by young children. Proceedings Interspeech, 749–752.