Recognition of foreign names spoken by native speakers

Frederik Stouten[1], *Jean-Pierre Martens*[1]

[1]ELIS, University of Ghent, Ghent, Belgium

fstouten@elis.ugent.be, martens@elis.ugent.be

Abstract

It is a challenge to develop a speech recognizer that can handle the kind of lexicons encountered in an automatic attendant or car navigation application. Such lexicons can contain several 100K entries, mainly proper names. Many of these names are of a foreign origin, and native speakers can pronounce them in different ways, ranging from a completely *nativized* to a completely *foreignized* pronunciation. In this paper we propose a method that tries to deal with the observed pronunciation variability by introducing the concept of a *foreignizable* phoneme, and by combining standard acoustic models with a phonologically inspired back-off acoustic model. The main advantage of the approach is that it does not require any foreign phoneme models nor foreign speech data. For the recognition of English names by means of Dutch acoustic models, we obtained a reduction of the word error rate by more than 10% relative.

Index Terms: spoken name recognition, phonological features, cross-lingualism

1. Introduction

It is a challenge to develop an automatic speech recognizer (ASR) that can accurately recognize proper names (e.g. person names, city names, street names, etc.) when the perplexity of the task is elevated. In a directory assistance application for instance, there may be a few 100K person names to distinguish. It is then extremely expensive to build a lexicon containing manually verified phonetic transcriptions. Hence, one must rely on an automatic grapheme-to-phoneme (G2P) converter instead. Such a G2P converter usually produces only one pronunciation per word. Now there is clear evidence (e.g. [1]) that, depending on their familiarity with the language of origin, native speakers may use different pronunciations of a foreign name, ranging from a totally *nativized* pronunciation (using native phonemes and native g2p rules) to a totally *foreignized* pronunciation (using foreign phonemes and foreign g2p rules). We therefore argue that the ASR should incorporate lexical and acoustic models that can cope with this type of pronunciation variability.

In [2] one proposes to use multiple G2P's to produce multiple pronunciations of a name: one G2P for the native language and one for each likely language of origin of the name (determined by running a language identification algorithm on it). The outputs of the non-native G2P's are then converted to a native phoneme sequence that is compatible with the native ASR. Adding pronunciations originating from foreign G2P's to the baseline dictionary caused a reduction of the word error rate (WER) by 25% for foreign names spoken by foreign speakers and by 10 % for foreign names spoken by native speakers.

In [3], one creates pronunciation variants in a data-driven way. This is achieved by using native acoustic models to align each name utterance with a graph of available initial pronunciations of that name (6 per name), and by seeking alternative phonemes for modeling regions where the acoustics badly match with the original phonemes. The newly created pronunciations induced an improvement of 20 to 40% over the initial pronunciations. However, the tests were run on the same names that were used to learn the new pronunciations.

A number of authors argue that in order to perform well, the ASR should incorporate acoustic models for non-native phonemes. In [4] for instance, models of English phonemes that have no good equivalent in German, are added to the inventory of acoustic models. These models are trained on a corpus of English speech provided by German speakers. By not converting the modeled foreign phonemes to German in the foreign manual transcriptions, one can achieve a reduction of the WER of 16.5% relative. In [5], non-native pronunciation variants are generated in a data-driven way. An English phoneme recognizer generates English pronunciations, and from an alignment of these pronunciations with the canonical pronunciations of the words one trains decision trees that can generate English-accented variants from German canonical transcription. This method yields a 5.2 % relative improvement.

In the last two approaches one needs foreign phoneme models. Consequently, if names from different languages are involved, one needs models for each of these languages. This may turn out to be impractical, especially when less-resourced languages are involved, like e.g. Indonesian or Russian. Therefore we propose a methodology that completely circumvents the need for foreign language acoustic models, and thus, for speech data from which to create such models.

Our proposal is to introduce *foreignizable* phonemes as native phonemes that can be foreignized to an attached foreign phoneme. The acoustic score for such a phoneme is obtained by combining the native acoustic model and a phonologically inspired back-off acoustic model that takes the properties of the attached foreign phoneme into account. The basic hypothesis is that foreign sounds can be represented by a set of phonological features, and that phonological feature models learned on native speech sounds are also able to characterize foreign sounds [6]. Obviously, the methodology can also be applied with phonological feature models that were trained on multilingual data (it is shown in [7] that these models are more reliable than monolingually trained models). We did not do that yet.

The outline of this paper is as follows. Section 2 explains our methodology: how to build the phonologically inspired back-off acoustic model, how to merge its scores with the classical acoustic model scores and how to introduce foreignizable phonemes in the lexicon. Section 3 describes the database and the spoken name recognition tests we conducted. Section 4 summarizes the most important conclusions of our work.

2. Methodology

Suppose that q represents a state of a baseline acoustic model, and that $\log p_A(\mathbf{x}|q)$ is the log-likelihood of acoustic vector **x** in this state. Then we propose to replace the baseline acoustic model score by a two-stream 'log-likelihood' score

$$LL(\mathbf{x}|q) = g_{1q} \log p_A(\mathbf{x}|q) + g_{2q}\, [\alpha\, \log p_B(\mathbf{x}|q) - \beta] \quad (1)$$

with $\log p_B(\mathbf{x}|q)$ representing the log-likelihood computed by means of a phonologically inspired back-off model, g_{1q} and g_{2q} the **state dependent** stream weights, and α, β the coefficients of a global linear model that aims at creating a score with the same over-all mean and variance as the baseline score. By introducing (α, β) we assume that the stream weights correspond to stream importances, meaning that we can restrict ourselves to $g_{1q} + g_{2q} = 1$.

2.1. Phonological feature models

In a previous paper [8] we introduced a phonological feature set of 25 binary phonological features (PHFs) to characterize acoustic-phonetic units. These features are denoted as f_i ($i = 1, .., 25$) and are grouped in four feature subsets: (1) **vocal source** (voiced, inactive), (2) **manner** (closure, vowel, fricative, burst, nasal, approximant, lateral, silence), (3) **place-consonant** (labial, labio-dental, dental, alveolar, post-alveolar, velar, glottal) and (4) **vowel-features** (low, mid-low, mid-high, high, back, mid, front, retroflex, rounded). Posterior probabilities $P(f_i|\mathbf{x})$ are estimated by a configuration of four neural networks (see [8] for more details).

2.2. Computing phonological scores

In order to determine $p_B(\mathbf{x}|q)$ we need to characterize each state q of a baseline HMM by its phonological features. For most phonemes, all states of the phoneme inherit the phonological features of this phoneme. However, some phonemes like plosives for instance, are modeled in terms of two acoustic-phonetic units with different phonological feature sets. The state q of such a phoneme then takes the phonological feature set of the acoustic-phonetic unit that best explains the acoustic observations assigned to this state during alignments of the training utterances with their orthographic transcription using the baseline acoustic models.

Since the phonological feature models compute posterior probabilities, log-likelihoods will be obtained as

$$\log p_B(\mathbf{x}|q) = \log \frac{P_B(q|\mathbf{x})}{P_B(q)} + \log p(\mathbf{x}) \quad (2)$$

where the subscript B indicates that these are probabilities according to the phonological model. Substituting this in Equation (1) leads to

$$\begin{aligned} LL(\mathbf{x}|q) &= g_{1q}\, \log p_A(\mathbf{x}|q) + \alpha\, g_{2q}\, \log p(\mathbf{x}) \\ &+ g_{2q}\, [\alpha\, \log \frac{P(q|\mathbf{x})}{P(q)} - \beta] \end{aligned}$$

We now assume that the second term is much less dependent on q than the other terms, and we use

$$LL(\mathbf{x}|q) = g_{1q} \log p_A(\mathbf{x}|q) + g_{2q}[\alpha\, \log \frac{P_B(q|\mathbf{x})}{P_B(q)} - \beta] \quad (3)$$

as the two-stream score.

Given the phonological description of q, the feature set can be divided in two subsets: P_q = the set of *positive* features that are supposed to be *on*, and N_q = the set of negative features that are supposed to be *off* for that state. Assuming independent phonological features then leads to the following expression:

$$\begin{aligned} \log \frac{P_B(q|\mathbf{x})}{P_B(q)} &= \sum_{f_i \in P_q} \log \frac{P(f_i|\mathbf{x})}{P(f_i)} \\ &+ \sum_{f_i \in N_q} \log \frac{1 - P(f_i|\mathbf{x})}{1 - P(f_i)} \end{aligned}$$

However, a statistical analysis of real data has shown that the two components in the right hand side of is expression are correlated (correlation coefficient of 0.75), and therefore, that it also makes sense to consider them as two estimations of the same log-likelihood ratio. Therefore, we propose to use some kind of means of the two as the ultimate estimator:

$$\begin{aligned} \log \frac{P_B(q|\mathbf{x})}{P_B(q)} &= w_{qp} \sum_{f_i \in P_q} \log \frac{P(f_i|\mathbf{x})}{P(f_i)} \\ &+ w_{qn} \sum_{f_i \in N_q} \log \frac{1 - P(f_i|\mathbf{x})}{1 - P(f_i)} \quad (4) \end{aligned}$$

We will investigate in particular what happens if (1) only positive or negative features are retained, and (2) not the mean of the log-likelihood ratios ($w_{qp} = w_{qn} = 0.5$) but the mean of the log-likelihood ratios per feature ($w_{qp} =$ 1/card(P_q) and $w_{qn} =$ 1/card(N_q)) are computed.

2.3. Foreignizable phonemes

The *baseline pronunciation* of a foreign name is normally obtained (see experiments) from its foreign transcription by mapping all foreign phonemes to their best equivalent in the native phoneme inventory. However, if this equivalent has another phonological feature representation than the original phoneme, we assume that it can be pronounced in a foreign way. We can introduce *alternative pronunciations* containing such so-called *foreignizable* phonemes. They appear with a foreign phoneme extension (see Table 1) in the lexicon. For the case of Dutch as the native and English as the foreign language, there are six English phonemes that have no equivalent with the same phonological representation in Dutch (see Table 1). If an English name then contains an /rr/, it will be mapped to /r_rr/, expressing that the normal pronunciation is /r/ but it can be pronounced in a foreign way as /rr/. When a foreign phoneme (e.g. /3:/)

Table 1: *English phonemes (SAMPA notation except for /rr/ and /r/: see www.phon.ucl.ac.uk/home/sampa) for which the Dutch equivalent has a different phonological representation. The representation differences are added in columns 3 and 4.*

Eng.	Du.	English PHFs	Dutch PHFs
Q	A	–	round
V	@	mid-low, back	mid-high, mid

3:	Y r	mid-low, mid	mid-high(Y), front(Y)
			round(Y), trill(r), alveolar(r)

aI	A j	–	round(A)

@U	O w	mid-high, mid	mid-low(O), back(O)
			round(O)

rr	r	approximant	trill, alveolar

is mapped onto a sequence of native phonemes (e.g. /Y r/), each of the latter is foreignized to that foreign phoneme. We consider all alternative pronunciations that can be obtained by substituting one or more foreignizable phonemes by their native equivalent. Table 2 shows two names and the variants created for them.

Table 2: *Two English names with their baseline and alternative native transcriptions. '_' represents a short pause.*

name	transcription	
Burr Tuppel	baseline	b Y r _ t Y p @ l
	alternative	b Y_3: r_3: _ t Y p @ l
Alan Presser	baseline	E l @ n _ p r E s @ r
	alternative 1	E l @ n _ p r_rr E s @ r
	alternative 2	E l @ n _ p r E s @ r_rr
	alternative 3	E l @ n _ p r_rr E s @ r_rr

2.4. Determining the stream weights

In all phoneme states we use the two-stream LL-score to assess the acoustic match of **x** to q. The stream weights are the same for all states of the same baseline acoustic model.

For a **native phoneme** state, the back-off model is called with the phonological description of the state as derived from the native phoneme. The stream weights are phoneme independent and presumed to be close to $g_1 = 1$ end $g_2 = 0$.

For a **foreignizable phoneme** state, the back-off model is called with the phonological description of the attached foreign phoneme. The stream weights are supposed to depend on the identity of the foreign phoneme, and they will be optimized experimentally (see next section).

3. Experiments

Our experiments are restricted to the recognition of English names by means of a Dutch speech recognizer. The acoustic models are 3-state triphone models. They were trained on CoGeN, a multi-speaker read speech database [9] capturing the Flemish variant of Dutch.

The English names were extracted from a spoken name database provided by Nuance Communications. We extracted 2050 name utterances: 21 different person names (first name + surname) spoken 3 or 4 times each (on different occasions) by 26 native speakers of Dutch. In order to raise the perplexity of the task, a lexicon of 1600 names was constructed: the 21 English names supplemented by 1579 Dutch person names. Automatically generated transcriptions of the English names were produced by the general-purpose Dutch and American English G2P converters from Nuance, and by a dedicated Dutch G2P converter that was trained on person names. The latter also generated two pronunciation variants per name [10]. Manual transcriptions of the English names were available too. Transcriptions of the Dutch person names were always generated with the general-purpose Dutch G2P.

3.1. Setting up a baseline system

First we investigated the effects of using different types of transcriptions for the foreign names in case the ASR works with baseline acoustic models only (no back-off model). The lexicons are named as follows:

DuAlone	foreign name transcriptions by Dutch G2P
DuMan	manual transcriptions added to DuAlone
EngAlone	foreign name transcriptions by English G2P
EngMan	manual transcriptions added to EngAlone
EngDu	foreign name transcriptions by English and Dutch G2P
EngVars	foreign name transcriptions by English G2P and name-specific Dutch G2P (2 variants per name)
ManAlone	only manual transcriptions of foreign names

The word error rates (WERs) and their 95% confidence intervals are listed in Table 3. The most important finding is that English transcriptions are very effective. When used alone, they even outperform (be it not significantly) the manual transcriptions. A lot of native speakers do seem to use a pronunciation that is closer to the foreignized than to the nativized pronunciation. A second finding is that manual transcriptions help a lot in combination with Dutch baseline transcriptions, but a lot less in combination with English transcriptions. This is because the manual transcriptions, when differing significantly from the baseline transciptions, are usually English-like transcriptions. Finally, the name specific G2P with variants does not outperform the general purpose G2P. This is probably because the former G2P was trained on a database comprising only a small fraction of English names.

Table 3: *Baseline performances (WER + 95% confidence intervals) for ASRs with baseline acoustic models but different pronunciation lexicons.*

lexicon	WER (%)	CI95 (%)
DuAlone	30.3	28.4 - 32.3
DuMan	23.5	21.6 - 25.3
EngAlone	23.1	21.2 - 24.9
EngDu	18.2	16.5 - 19.9
EngMan	16.8	15.2 - 18.4
EngVars	18.1	16.5 - 19.8
ManAlone	24.7	22.8 - 26.5

3.2. Testing the proposed methodology

Since one usually has no access to manual transcriptions or a name-specific G2P, we take lexicon *EngDu* as our baseline lexicon and we create pronunciation variants from the English transcriptions by means of the procedure outlined in Section 2.3.

We first test the phonological back-off model as the only acoustic model in the ASR. If we take both positive and negative features into account, the WER is 41.4% when averaging unnormalized log-likelihood ratios and 37.2% when averaging normalized log-likelihood ratios. Omitting the negative features in the two situations, pushes the WER to 43.6% and 38.3% respectively. Consequently, we will use all phonological features and normalized log-likelihoods.

The next step is to determine the optimal stream weights for a particular foreignizable phoneme. We do that by removing all variants containing other foreignizable phonemes, and by performing a recognition test with the retained variants for four different values of g_1. We then select the g_1 yielding the lowest WER as the stream weight of this phoneme. We repeat the whole process until we have appropriate weights for all foreignizable phonemes. Table 4 shows the recognition results for

Table 4: *Effects of the stream weights on the WER of an ASR incorporating a phonologically inspired back-off model. For each foreign phoneme we included how many times it appears in a transcription (all transcriptions together count 304 phonemes).*

phoneme	g_1	g_2	WER (%)
/3:/ (1)	0.6	0.4	17.7
	0.4	0.6	17.8
	0.2	0.8	17.9
/Q/ (14)	0.6	0.4	17.9
	0.4	0.6	17.8
	0.2	0.8	17.7
/V/ (2)	0.8	0.2	18.2
	0.6	0.4	18.1
	0.4	0.6	18.1
/aI/ (1)	0.8	0.2	18.1
	0.6	0.4	18.1
	0.4	0.6	18.4
/@U/ (2)	0.8	0.2	18.1
	0.6	0.4	18.1
	0.4	0.6	18.1
/rr/ (22)	0.4	0.6	17.8
	0.2	0.8	17.4
	0.1	0.9	17.7
all	opt.	opt.	**16.5**
native	0.8	0.2	17.4
	0.7	0.3	17.3
	0.6	0.4	17.8
all+native	0.7	0.3	**16.2**

the six foreignizable phonemes we investigated. Apparently, the positive effect of the back-off model is significant for phonemes that occur frequently in the transcriptions (/rr/ and /Q/), and for /3:/, even though it appears only once in the lexicon. If optimal stream weights are used for **all** foreignizable phonemes (label *all*), then the WER drops to 16.5%, which represents an improvement 9.3% relative over the baseline system.

By also using optimized phoneme independent stream weights for the native phonemes (label *native*) in the ASR, we obtain a WER of 16.2% (label *all+native*) representing an improvement over the baseline system of 11% relative. The optimal stream weights for the native phonemes clearly favor the standard acoustic model whereas the opposite is true for the foreignizable phonemes.

We also tested the method in combination with a lexicon having only two pronunciations per name: the baseline pronunciation and the pronunciation with all the foreignizable phonemes present. Using the same stream weights as before, the WER now becomes 17.1%. Apparently, it is better to let the ASR choose between a small and a large importance of the back-off model in the LL-scores of foreignizable phonemes.

4. Conclusions

We have proposed a technique for improving the recognition of foreign names spoken by native speakers. The method is based on the introduction of foreignizable phonemes and two-stream acoustic models for these phonemes. The two-stream models combine the standard acoustic likelihood on a triphone state with a phonological score for that state. The phonological score is derived from posterior phonological feature probabilities and from the phonological representation of a foreign phoneme that is associated with the native phoneme. The posterior phonological feature probabilities are computed by means of neural networks that were trained on native speech only. The latter means that the presented method does not require any foreign phoneme models, nor a speech corpus containing foreign phonemes from which to train foreign pronunciations.

For the recognition of English person names spoken by Dutch speakers, the accuracy of our Dutch ASR was improved by 11% relative. This comes on top of the 40% that was obtained by including baseline pronunciations derived from an English G2P. The improvement is achieved with a small additional cost, originating from the computation of phonological scores and the inclusion of extra variants in the lexicon.

We are aware that our small-scale experiment only offers a proof of concept and that a test on a larger database with more different foreign names and more foreign languages is in order. We are currently preparing such a test (600 different names, three foreign languages: English, French, Moroccan).

5. Acknowledgements

This work was supported by the Flemish Institute for the Promotion of Scientific and Technical Research in the Industry (contract SBO/40102). Nuance is acknowledged for making available the G2P converters and the name corpus.

6. References

[1] S. Fitt, "The pronunciation of unfamiliar native and non-native town names," in *Proc. Eurospeech*, Madrid, Spain, 1995, pp. 2227–2230.

[2] B. Maison, S.F. Chen, and P.S. Cohen, "Pronunciation modeling for names of foreign origin," in *Proc. ASRU*, Virgin Islands, USA, 2003, pp. 429–434.

[3] F. Beaufays, A. Sankar, S. Williams, and M. Weintraub, "Learning name pronunciations in automatic speech recognition systems," in *International Conference on Tools with Artificial Intelligence*, 2003, pp. 233–240.

[4] G. Stemmer, E. Nöth, and H. Niemann, "Acoustic modeling of foreign words in a german speech recognition system," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 2745–2748.

[5] S. Goronzy, S. Rapp, and R. Kompe, "Generating non-native pronunciation variants for lexicon adaptation," in *Speech Communication*, 2004, vol. 42, pp. 109–123.

[6] G. Williams, M. Terry, and J. Kaye, "Phonological elements as a basis for language-independent asr," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 88–91.

[7] S. Stüker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in *Proc. ICASSP*, Hong Kong, China, 2003, pp. 144–147.

[8] F Stouten and J.-P. Martens, "On the use of phonological features for pronunciation scoring," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 229–232.

[9] K. Demuynck, D. Van Compernolle, C. Van Hove, and J.-P. Martens, "Een corpus gesproken nederlands voor spraaktechnologisch onderzoek," in *Technical Report ESAT - ELIS*, 1997, pp. 1–30.

[10] Q. Yang and J.-P. Martens, "Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names," in *Proc. LREC*, 2006, pp. 287–292.