ISCA Archive
http://www.isca-speech.org/archive

ITRW on Adaptation Methods
for Speech Recognitionn
Sophia Antipolis, France
August 29-30, 2001

Improving the recognition of foreign names and non-native speech by combining multiple grapheme-to-phoneme converters

Nick Cremelie & Louis ten Bosch

Lernout & Hauspie Speech Products nv
Koning Albert I laan 64, B-1780 Wemmel, Belgium
{nick.cremelie,louis.tenbosch}@lhs.be

Abstract

Tasks involving the recognition of native as well as foreign names and being accessed by native as well as non-native speakers are not handled very well by a baseline ASR system with single language acoustical models and one transcription per word in its lexicon. One possible solution is lexical adaptation: alternative transcriptions allowing for a closer match with actual pronunciations from the speakers, are added to the lexicon. In this paper it is shown how transcriptions created by foreign language grapheme-to-phoneme converters can be integrated adequately in the recognizer and serve the objected purpose. The experiments reported here show that this approach can effectively deal with foreign names and non-native speakers, as WER reductions of almost 60% relative are attained.

1. Introduction

An important number of recognition tasks involves the recognition of names: automated attendant or name dialing tasks (proper name recognition), navigational query tasks (location name recognition) [1], etc. Within these tasks one has to deal with some specific lexical problems:

- **Wrong or bad transcriptions:** since names are most often not available in a pronunciation dictionary, the recognizer has to fall back to an automatic grapheme-to-phoneme (G-to-P) converter in order to retrieve a phonetic transcription of the names. However, as the vocabularies of the tasks mentioned may comprise foreign names, very often a faulty transcription is generated. These can severely hurt the recognizer's performance.
- **Non-native speech:** the applications mentioned are commonly used by native as well as non-native speakers. Depending on their proficiency level for the target language, the non-native speakers may use phonemes from their own native language instead of the correct ones to pronounce target language names, or even completely pronounce the names in a way consistent with their native language.

Although the recognition of non-native speech may require a combination of several adaptation techniques (ranging from classical speaker adaptation to multilingual acoustical modeling [2,3,4,5]), we have investigated how far a purely lexical adaptation approach can bring us in the context of name recognition tasks.

We do not resort to a data-based lexical adaptation method. These have proven successful in tuning lexicons for general recognition tasks (see e.g. [6]), but they require an offline training on a suitable training database and mostly concentrate on general pronunciation variation mechanisms. The tasks we are looking at here exhibit specific pr1oblems related to multilinguality, leading to low baseline recognition performance. Therefore we have searched for a solution that raises the performance significantly and can be applied to any new vocabulary without the need of collecting training data.

The starting point is a recognizer trained for one particular language and a lexicon with just one transcription per word, created by a native language G-to-P converter. From there, the lexicon is augmented with new transcriptions generated by foreign language G-to-P converters. A G-to-P converter typically relies on a set of pronunciation rules in order to determine a phonetic transcription for a given orthographical word. G-to-P converters are commonly used in the context of text-to-speech systems, but prove useful as well within ASR systems.

In the next section, we will first show how new G-to-P transcriptions can be integrated in the recognition system. Section 3 will discuss some experiments with the outlined approach. In section 4, we compare the single recognizer / multiple transcriptions solution with a multi-recognizer approach. Finally, conclusions are summarized in section 5.

2. Integrating Foreign G-to-P Transcriptions

2.1. Basic principle

As stated before, the baseline lexicon contains one transcription per name, created by a G-to-P converter for the target or native language (i.e. the language for which the recognizer was trained). Next, alternative transcriptions for the names are generated by G-to-P converters for foreign languages. The correct (apart from possible G-to-P conversion errors) phonetic transcription of a name is obtained when the language of the G-to-P converter matches the language origin of the name. Yet, it makes sense to keep all alternative transcriptions of each name in the lexicon, for various reasons:

- The language origin of each name is normally not given, and hence it is not possible to identify the correct transcription in an automatic way.
- Some names are common in different languages and may have different pronunciations in each of these languages.
- The speaker — whether be it a native or a non-native — may not know how to pronounce a foreign name (i.e. originating from a language different from his native language). In such case, he will usually try to pronounce

that name in a way consistent with his native language (or sometimes even with a second language he knows). This pronunciation will usually correspond pretty well with the transcription of that name produced by a G-to-P converter for the speaker's native (or second) language. Hence, by letting all G-to-P converters transcribe all names, regardless of their language origin, we automatically include transcriptions reflecting how speakers pronounce names that are foreign to them. Of course, this only holds for speakers with a low proficiency level in the foreign language, as more proficient speakers may have a clue about how to pronounce the foreign name at least partially correct. Still, the non-native G-to-P transcriptions may represent a closer match with actual pronunciations used by those speakers.

2.2. Phoneme mappings

One must realize however that adding foreign G-to-P transcriptions to the baseline lexicon is not as straightforward as it may seem. Since the goal is to use a standard recognizer with acoustical models trained for only one language (the native language), we are limited to the phoneme set of these models. However, the foreign transcriptions may be described in terms of a different phoneme set. As such, they may contain phonemes not available in the acoustical model set.

This problem can be overcome by first *mapping* the foreign phoneme symbols onto native phoneme symbols. We choose to base these mappings on the phonetic knowledge present in the IPA phonetic alphabet definition [7]. The IPA alphabet has symbols for phonemes in any language, and within the definition of the symbols it is indicated how close to or how distant from each other the phonemes are, in terms of features such as manner or place of articulation. Such mapping will of course not be perfect, as a good equivalent native phoneme for a certain foreign phoneme may simply not exist. However, we found from our experiments that these mapping can be very effective if we do not stick to a one-to-one mapping but rather introduce two additional possibilities:

- Allow a single phoneme to be mapped onto a sequence of phonemes; e.g. map English /aɪ/ to Dutch /a j/. This allows some phonemes to be mapped more accurately.
- Allow a single phoneme to be mapped onto more than one phoneme, thus providing alternatives. This is useful when more than one alternative lies close to the original phoneme. E.g. map English /æ/ to Dutch /a/ as well as /ɛ/.

Both mechanisms can be combined as well. Creating these mappings obviously requires some expertise, but basically, with the IPA as a guidance, it is not an impossible job to find a decent mapping for not too distantly related languages.

2.3. Pre- and post-processing

Special care needs to be taken when names are given to a foreign language G-to-P converter.

First of all, the name may contain letters or other symbols that the foreign G-to-P converter does not understand. E.g. while accented letters are very common in French, they are basically not used in English; hence, an English G-to-P converter might not know what to do with the accented letters present in a French name. Therefore, the names generally need some pre-processing before they are sent to the foreign G-to-P. Referring to the previous example, one could replace the accented letters by their non-accented counterparts.

Secondly, after careful investigation of G-to-P outputs on foreign names, we found that some mild post-processing on the G-to-P output can be beneficial as well. A G-to-P converter usually converts graphemes into phonemes through some pronunciation rules. When confronted with a foreign name, it blindly applies its rules as if it was a word from its own language. In our case, a statistical analysis reveals that the G-to-P outputs on foreign names often contain unusual but systematically reappearing phoneme pairs. Hence, we identify the most common of those "mistakes" and correct them by applying appropriate rewrite rules. This form of post-processing results in more consistent transcriptions.

2.4. Penalty weights

Another issue related to the integration of multiple transcriptions per word in the lexicon are *penalty weights*. Each transcription of a word can be assigned a penalty weight in order to discriminate it amongst its siblings. There is clear evidence in the literature that such weights (or transcription probabilities) lead to a gain in recognition accuracy [6], although the gain seems to depend on many factors, like e.g. the ASR engine design, the accuracy of the acoustical models, the accuracy of the weight estimations, the task at hand, etc. The best approach is to train these weights on a set of data (perform recognition on the data and make a statistical analysis of how often each transcription was actually taken). However, in many cases an appropriate training set is not available, or there is simply no opportunity to perform such a training prior to engaging the recognition system. Therefore we explore two alternatives here.

The first one is to assign *global* penalty weights to the transcriptions. All transcriptions produced by a particular G-to-P converter receive the same penalty weight. This means that all transcriptions in the same language have the same weight. These weights are set a priori.

The second alternative is to rely on a so-called *language identifier*. This tool estimates, for a given word and based on its orthography, the probabilities of the different possible language origins of that word. From these probabilities one can compute the desired transcription weights. The weight w_l for a G-to-P transcription of language l for a certain word x is determined as:

$$w_l = M \cdot \left[1 - \frac{\Pr_l}{\Pr_{\max}}\right] \qquad (1)$$

where $\Pr_l$ is the language identifier probability for language l on word x (i.e. Pr(language origin of x is l | language identifier)), and $\Pr_{\max}$ is the largest language identifier probability obtained on word x. The factor M is a scaling factor that sets the upper boundary on the range of the weights. It can be optimized in the same way a general word penalty is optimized for a certain task.

3. Experiments & Results

3.1. Test material and environment

We have run a series of experiments with two different databases:

- **NAM-1:** this database contains loggings from an automated attendant system in operation. The recordings are telephone quality (8kHz). The lexicon of the automated attendant task comprises 406 proper names. During recognition, a grammar defining the valid combinations of first names and surnames is used. The test set contains 1000 utterances. The database has been validated and the speakers are tagged according to their native language. About 75% of the speakers has a Dutch background, 10% are native English speakers and 5% are native French speakers. The remaining 10% is identified as Other/Unknown.

- **NAM-2:** this database contains name dialing instructions. The recordings are made in an office environment, sampling frequency is 11kHz. The lexicon consists of 230 names. Again, a grammar links first names to surnames. The test set contains 3000 utterances. The speakers are not identified according to their native languages, but it is known that Dutch, English and French are the major language backgrounds.

Recognition experiments are performed with an HMM-based continuous speech recognizer. Since Dutch is the (or a) major language background of the speakers in both databases, Dutch acoustical models are plugged into the recognizer. For NAM-1, we use telephony grade models, while for NAM-2 more accurate office grade models are used.

3.2. Experiments on NAM-1

We consider several lexical configurations and run recognition experiments with each of those:

1. *D-Baseline*: each name in the lexicon is transcribed by the Dutch G-to-P converter only.
2. *DFE-Multi*: French and English G-to-P transcriptions are added to the baseline lexicon. As explained before, the foreign phonemes within these transcriptions are mapped onto the native phoneme set — in this case the Dutch set, as it is known by the Dutch acoustical models. Note that, because we allow multiple phoneme mappings, each foreign transcription containing one or more of these multiple-mapped phonemes will translate into multiple transcriptions.
3. *DFE-Multi+WP*: same as 2., but additionally a word-independent word penalty (i.e. one system-wide parameter, attached to each word) is introduced and optimized.
4. *DFE-Multi+GlobalWeights*: same as 2., but additionally a penalty weight is attached to each transcription. These weights are language-global, meaning that all transcriptions coming from the same G-to-P receive the same weight.
5. *DFE-Multi+LangIdentWeights*: same as 2., but the transcriptions now receive individual penalty weights derived from the language identifier probabilities.
6. *DFE-Manual*: the original baseline lexicon is hand-optimized by linguists. Faulty G-to-P transcriptions are corrected, foreign names are transcribed correctly, and alternative transcriptions reflecting how native Dutch, French or English speakers would pronounce names that are foreign to them are included. This can be considered as *the best possible lexicon tuning*.

The word error rates (WER) emerging from recognition experiments on NAM-1 with each of the above lexicon configurations are summarized in Table 1.

Lexicon Configuration	WER
D-Baseline	20.0%
DFE-Multi	16.8%
DFE-Multi+WP	15.0%
DFE-Multi+GlobalWeights	12.8%
DFE-Multi+LangIdentWeights	14.3%
DFE-Manual	10.4%

Table 1: WERs on NAM-1 test set with different lexicon configurations.

The results lead to some interesting conclusions. First of all, the introduction of multiple foreign G-to-P transcriptions in the lexicon does result in a significant reduction of the WER. Secondly, the lowest WER obtained with these transcriptions approaches the one with the manually optimized lexicon, indicating that the G-to-P transcriptions exploit almost the full potential of lexicon adaptation. Third, assigning penalty weights to the transcriptions seems to contribute importantly to the WER reduction, although an optimized word penalty can partially account for that reduction in the absence of penalty weights. The language identifier weights perform worse than the global weights. This can be explained as follows: the language identifier weights penalize transcriptions coming from a G-to-P converter whose language does not match the language origin of the word. E.g. if, according to the language identifier, there is only a small chance that a certain word is a French word, the French transcription of that word will receive a high penalty. These penalties actually work against non-native speakers who do use those transcriptions.

The effect of the foreign transcriptions is analyzed in more detail for *D-Baseline* vs. *DFE-Multi+GlobalWeights*. The WER is split out per speaker language background based on the speaker tags present in the database (see Table 2).

Speaker's Language Background	*D-Baseline*	*DFE-Multi-GlobalWeights*
Dutch	15.1%	9.3%
French	33.3%	10.3%
English	41.0%	22.9%
Other/Unknown	28.4%	24.7%

Table 2: WERs split out according to speaker's language background.

Clearly, the French and English speakers benefit to a large extent from the foreign transcriptions. More than that, the WER is reduced for the Dutch speakers as well, indicating that some of these speakers use the correct pronunciation of

some of the French and/or English names and, with the correct transcription of those names being present in the lexicon, they are subsequently recognized correctly.

Finally, we have verified to what extent the improvement for the French and English speakers was due to the inclusion of non-native transcriptions (i.e. French or English transcriptions of Dutch names). For that purpose, the names are tagged according to their language origin. That way, we find e.g. that the WER on Dutch names drops from 24.0% to 16.0% for the French speakers and from 48.9% to 29.8% for the English speakers after adding French and English G-to-P transcriptions of these names. Overall, roughly one third of the improvement can be attributed to non-native transcriptions for this task.

3.3. Experiments on NAM-2

On NAM-2 we run the same series of experiments as on NAM-1, expect that this time no DFE-Manual configuration is available. The results are displayed in Table 3.

Lexicon Configuration	WER
D-Baseline	16.4%
DFE-Multi	7.1%
DFE-Multi+WP	7.1%
DFE-Multi+GlobalWeights	7.0%
DFE-Multi+LangIdentWeights	6.7%

Table 3: WERs on NAM-2 test set with different lexicon configurations.

Again a considerable WER reduction appears. Notably, the configuration using weights from the language identifier now delivers the lowest WER while on NAM-1 (Table 1) the global weights are the top performers. Apparently, the gain from putting penalty weights on the transcriptions is much lower here. However, the results with the more realistic language identifier weights lead to the assumption that more accurate acoustical models require more accurate penalty weights. We tend to believe that weights trained on actual speech data would even further reduce the WER, however this would be a less flexible approach as it is not robust against the addition of new words or transcriptions.

4. Recognizers in Parallel

For comparison reasons, we set up a system with three recognizers (a Dutch, a French and an English one) placed in parallel. Each recognizer receives the complete lexicon, but only with the G-to-P transcriptions of the words for its own language. The N-best lists coming out of these recognizers are combined, based on weighted confidence scores, into a uniform recognizer output. In this way we obtain a multi-lingual system. This new system has acoustical models for phonemes from all three languages and therefore it is supposedly more accurate at the acoustical level. The major drawback is the additional memory and processing overhead required for running three recognizers in parallel.

A recognition experiment with this system is run on NAM-1. The purpose of this experiment is to check whether in this way higher accuracy rates can be attained than with a single recognizer / multiple transcriptions system. Apparently, this is not the case as the WER on NAM-1 amounts to 15.5% with this system, a good deal above the 12.8% obtained previously. Consequently, the baseline Dutch recognizer in the multi-recognizer system is replaced by the *DFE-Multi+GlobalWeights* configuration. This time, the WER on NAM-1 drops down to 11.6%.

We can conclude from these results that in the tasks at hand the acoustical models are not really the main issue. Although having acoustical models for the foreign language involved is helpful (as the latter result shows), much more is to be gained from a more adequate pronunciation modeling.

5. Conclusion & Final Remarks

We have shown how foreign G-to-P transcriptions can be integrated in an existing baseline system. The experiments proved this method to be extremely successful in reducing the WER for name recognition tasks accessed by native as well as non-native speakers. Some other interesting observations emerged from the experiments as well:

- Adding penalty weights to the transcriptions does improve the recognition accuracy. However, more accurate acoustical models must be combined with more accurate weight estimations.
- Adding recognizers for all the major language origins of the names and the speakers does not result in lower WERs than a single language system combined with multiple transcriptions.

As a final note, we point out that a smoother combination of both approaches tested here may very well lead to the best solution for the targeted tasks. Starting with a set of acoustical models for one language, augmenting it with models for the most crucial missing phonemes from the major foreign languages in the task, and finally adding the foreign G-to-P transcriptions to the lexicon is likely to robustify the system and to push down the WER even further.

6. References

[1] I. Trancoso, C. Viana, I. Mascarenhas, C. Teixeira (1999), "On deriving rules for nativised pronunciation in navigation queries", in Proceedings Eurospeech 1999.

[2] L. Tomokiyo (2000), "Lexical and acoustic modelling of non-native speech in LVCSR", in Proceedings of ICSLP2000.

[3] I. Amdal, F. Korkmazskiy, A. Surendran (2000), "Joint pronunciation modelling of non-native speakers using data-driven methods", in Proceedings of ICSLP2000.

[4] K. Livescu, J. Glass (2000), "Lexical modelling of non-native speech for ASR", in Proceedings ICASSP2000.

[5] U. Uebler, M. Boros (1999), "Recognition of non-native German speech with multilingual recognizers", in Proceedings Eurospeech '99.

[6] N. Cremelie, J.-P. Martens (1999), "In search of better pronunciation models for speech recognition", in Speech Communication, vol. 29.

[7] International Phonetic Association (1999), "Handbook of the International Phonetic Association", Cambridge University Press.