

Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration

Ghazi Bouselmi, Dominique Fohr, Irina Illina, Jean Paul Haton

Speech Group, “<http://parole.loria.fr/>”
LORIA, Nancy, France

{ bousselm, fohr, illina, jph }@loria.fr

Abstract

This paper presents a fully automated approach for the recognition of non-native speech based on acoustic model modification. For a native language (L1) and a spoken language (L2), pronunciation variants of the phones of L2 are automatically extracted from an existing non-native database as a confusion matrix with sequences of phones of L1. This is done using L1’s and L2’s ASR systems. This confusion concept deals with the problem of non existence of match between some L2 and L1 phones. The confusion matrix is then used to modify the acoustic models (HMMs) of L2 phones by integrating corresponding L1 phone models as alternative HMM paths. In this way, no lexicon modification is carried. The modified ASR system achieved an improvement between 32% and 40% (relative, L1=French and L2=English) in WER on the French non-native database used for testing.

1. Introduction

In the last twenty years, researches in automatic speech recognition has lead to huge advances. Recent ASR systems achieve high recognition rates. Nevertheless, the performance of these systems drops drastically when confronted with pronunciations that deviate from canonical lexicon definition, especially with non-native speakers. This drastic performance drop when handling non-native speech is a well known problem.

The main aim of non-native enhancement of ASRs is to make the available recognition systems tolerant to pronunciation variants. This is done by integrating some extra knowledge into existing systems. This extra knowledge corresponds to the pronunciation variants to be handled by the system: dialects, accents or non-native variants. Approaches differ in the techniques used to extract this knowledge and integrate it into an existing native ASR system. The next section spots some of the recent researches in non-native speech recognition. In these approaches, modifications are made in different layers of the outlying baseline systems, varying from lexicon ([1],[2] and [3]) to Gaussian mixture parameters ([4] and [5]), and HMM models ([6]).

In [1], knowledge about non-native speech accent is acquired through a study of phonological properties of both the spoken language and the native language of the speaker. The extracted knowledge is represented as a set of rewriting rules in which phones of the spoken language are replaced by phone of the native language. These rules are language pair specific (spoken/native) and are used to modify the lexicon of the spoken language ASR system.

In [2], phonetic confusion is automatically extracted from non-native speech database by aligning the canonical pronunciation of each utterance with its actual pronunciation (phonetic transcription of the utterance obtained after a phonetic recognition with the baseline spoken ASR system). This confusion is then used to modify the lexicon by dynamically adding all possible phonetic transcriptions of each word during the recognition phase.

In [4], both spoken and native language ASRs are used to extract the phonetic confusion. The native language ASR is used to obtain a phonetic transcription (in terms of native phones) of all non-native utterances. Confusion is extracted by aligning the latter transcription with the canonical one (obtained from the lexicon of the spoken language) for each utterance. According to this confusion, Gaussian mixture models of native phones are merged with Gaussian mixture models of the spoken language phones (for each state of the HMMs). These modified phone models are then used as new models for the spoken language ASR system.

2. Our new approach

The main motivation of our work is to develop a new approach for non-native speech recognition that can automatically handle non-native pronunciation variants without a significant loss in recognition performance. As non-native speakers tend to realize phones of the spoken language as they would do with similar phones from their native language, we claim that taking into account the acoustic models of the native language in the modified ASR system may enhance its performance.

For instance, the sound ‘[ð]’ (present in word *the*, using IPA alphabet) does not exist in French. A high percentage of French native speakers realize this phone as the French phone ‘[ʒ]’. Furthermore, diphthongs like ‘[tʃ]’ (present in word *church*) do not exist in French. The latter diphthong may however be uttered as the two French phones ‘[t] [ʃ]’ or ‘[tʃ]’ as stated by phonetician experts.

Thus, in our approach, the confusion involves a phone of the spoken language and a phone sequence of the native language. The main idea is to automatically extract a confusion between spoken language phones and sequences of phones of the native language using both language ASRs.

Besides, this confusion will be utilized by means of

HMM modification rather than by lexicon modification. As stated in [2], injecting confusion knowledge into the lexicon may result in an excessive growth of the lexicon and thus of the search space. The recognition would be very slow unless some pruning is used. Furthermore, merging the Gaussian mixture models of each state of the HMM of the confused spoken and native language phones (as in [4]) may deteriorate the coherence of the acoustic models of both phones.

In our approach, the confusion is extracted using the two time-aligned transcriptions given by the spoken and by the native language ASRs. The acoustic model (HMM) of each spoken language phone is modified by integrating the acoustic models (HMMs) of each native language phone sequence it was confused with. This process is described in the next sections.

2.1. Confusion extraction

Both spoken language and native language ASR systems are used for confusion extraction. For each utterance of the non-native speech database, a forced phonetic alignment is performed using the spoken language ASR system, and then a phonetic recognition is performed using the native language ASR system. This provides a time-aligned canonical phone transcription for the first system (in terms of spoken language phones) and a time-aligned actual phone transcription for the second system (in terms of native language phones). These two time-aligned transcriptions are then compared in order to detect the sequence of native phones that was recognized for each spoken language phone in the utterance. Given a spoken language phone L present in the utterance, the sequence associated with L is composed of native language phones whose time interval is included in L 's time interval. As phonetic recognition does not provide exact phone boundaries, the latter condition has to be slightly relaxed. To be taken into account in the associated phone sequence, a phone of the native language must have at least 50% of its time interval included in the underlying spoken language phone time interval. In the example of figure 1, the sequence of native language phones (M_1, M_2) would be associated with phone L because M_1 and M_2 have more than 50% of their time interval included in L 's one, and M_3 does not.

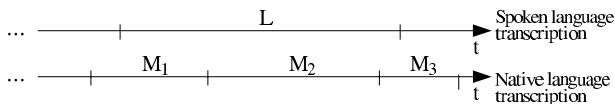


Figure 1: Example of time-aligned transcriptions (for the same utterance).

The next step is to extract the confusion rules from the above phone and phone sequence associations. Having the count of appearance of each association, the maximum likelihood (ML) estimate of the confusion probability is then computed as follows (for each spoken language phone L):

$$\begin{aligned} P(L \implies \{M_i\}_{i \in I}) &= P((M_i)_{i \in I} | L) \\ &= \frac{N(L \implies (M_i)_{i \in I})}{N(L)} \end{aligned} \quad (1)$$

where $N(L \implies (M_i)_{i \in I})$ is the count of appearance of the underlying association $L \implies (M_i)_{i \in I}$, I a set of indices, and $N(L)$ the count of appearance of the phone L .

Finally, only the confusion rules that have the highest probability (satisfying the condition in equation 2) are taken into account. This way, the use of erroneous rules that resulted from phonetic recognition errors is reduced. Besides, this thresholding avoids the excessive growth of the phone models that will be used later in the modified system.

$$\frac{P(L \implies (M_i)_{i \in I})}{\max_{x \in R_L} P(x)} \geq \alpha \quad (2)$$

where R_L is the set of rules having the phone L as left part, and α a threshold.

Here are some examples of the rules given by our system when run with English as spoken language and French as native one:

- for phone '[p]' (present in word *pet*):
"[p]==>[p]" $P([p]==>[p]) = 1$
- for phone '[r]' (present in word *absorb*):
"[r]==>[l]" $P([r]==>[l]) = 1$
The English phone '[r]' does not exist in French. Rather, its closest French phone is '[l]'.
- for diphthong '[tʃ]' (present in word *church*):
"[tʃ]==>[t][ʃ]" $P([tʃ]==>[t][ʃ]) = 0.443$
"[tʃ]==>[k][ʃ]" $P([tʃ]==>[k][ʃ]) = 0.286$
"[tʃ]==>[f]" $P([tʃ]==>[f]) = 0.271$

2.2. HMM integration

The second part of our approach consists in applying the confusion rules to the ASR system. We use a novel method in this step that integrates all HMM corresponding to the confused phone sequences into the HMM model of the underlying spoken language phone. Figure 2 illustrates the HMM structure used in our ASR system for each phone. These models have three emitting states all linked in a left-to-right path.

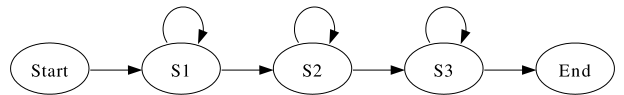


Figure 2: Phone HMM model structure.

The modification of the HMM of a spoken language phone L consists in adding new state paths into it. Each of these added state paths corresponds to one of the confusion rules of R'_L (R'_L is the selected rules according to the previous section, $R'_L \subseteq R_L$). Each of them is constructed by the concatenation of the HMM models of each phone present in the left part of the underlying rule.

The transition linking the *Start* state to the state $S1$ of the spoken language phone has a probability of β . Here β is the weight of the original spoken language model versus the models introduced by the confusion. The transition linking the *Start* state to each HMM path representing a rule $r \in R'_L$ has a probability $P'(r)$:

$$P'(r) = (1 - \beta) \frac{P(r)}{\sum_{x \in R'_L} P(x)} \quad (3)$$

Assuming the rules sketched in section 2.1, figure 3 illustrates the construction of the modified HMM for the English phone [tʃ].

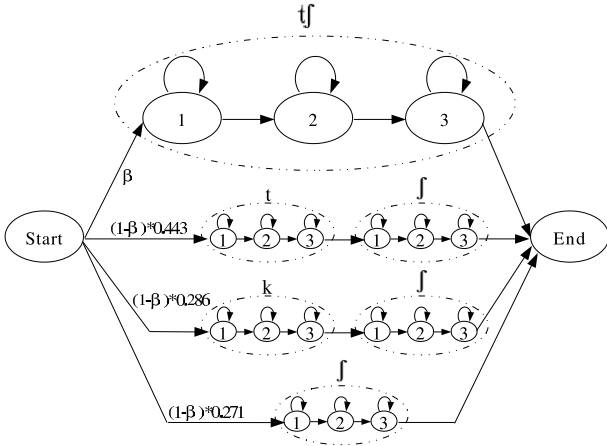


Figure 3: Modified HMM model structure for English phone [tʃ].

3. Experiments

The work presented in this paper has been done in the framework of the European project *HIWIRE* which aims at enhancing speech recognition in mobile, open and noisy environments. Actually, the *HIWIRE* project deals with the development of an automatic system for the control of aircrafts by pilots via voice commands.

Pilots have to speak in English regardless of their native language. Thus, the major part of speech is non-native. For now, we worked only on French speakers speaking English.

3.1. Experimental conditions

The used acoustic parameters are 13 MFCCs with their first and second time derivatives. The 46 English monophone models have been trained on the *TIMIT* database which contains 420 speakers and 3360 utterances (in its training set). The 40 French monophone models have been trained on the French database *ESTER* which contains 90 hours of broadcast news. The HMM models used in the two base line ASR systems (French and English) have 128 Gaussian mixtures per state and diagonal covariance matrices.

The non-native database contains 21 French speakers with 100 utterances for each. It was recorded at a sampling rate of 16Khz at 16 bits per sample. Half of this database was used for development, the other half for testing.

The vocabulary is composed of 134 words, and the grammar is a command language. We also used a second non-restrictive grammar in our tests, i.e., a “word-loop grammar” that allows the recognizer to choose any sequence of words present in the lexicon.

3.2. Confusion extraction issues

As described above, the confusion is extracted by comparing a transcription given by phonetic alignment (English ASR) with a transcription given by a phonetic recognition (French ASR). The French phonetic recognizer was tuned to give the best value of the ratio:

$$R_{phone} = \frac{N_{French}}{N_{English}} \quad (4)$$

where $N_{English}$ is the count of English phones in all utterances and N_{French} is the count of French phones given by

the French phonetic recognizer. Based on the count of simple phones and diphthongs in English, we computed a value of $R_{phones} = 1.2$.

3.3. Development and results

We used a value of β equal to 0.5 for all further tests (see 2.2) in order to give the same weight to both spoken and native language acoustic models. As for the confusion extraction, the modified English ASR system had to be tuned in order to have the best results. We used the “word accuracy (WACC)” criterion to find the best word-insertion-log-penalty. Figure 4 shows the variation of the “DEL”, “INS” and “WACC” versus the value of word-insertion-log-penalty. The best value for the penalty is -50. This value was used for all further tests.

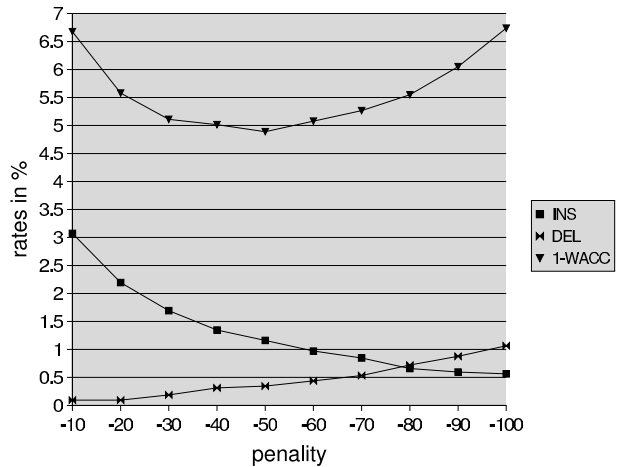


Figure 4: Rates versus penalty (modified system, development database).

In the rest of this section, we will use the following notation:

- “baseline system” refers to the spoken language ASR system, with English phone models trained on the *TIMIT* database (see 3.1).
- “fully automated confusion (FAC)” refers to the ASR proposed in this paper with modified phone models (according to sections 2.1 and 2.2).
- “*TIMIT* confusion” refers to the “fully automated confusion” (the latter system) except that the confusion was extracted using *TIMIT* database instead of the non-native database.
- “expert confusion” refers to the “fully automated confusion” except that we used a “confusion” given by a phonetician expert. Here, the expert associated English phones with nearest French ones in terms of acoustics. The expert did not take into account any non-native variants or pronunciation errors.

We tested both the baseline and the “fully automated confusion” systems with the grammar presented in 3.1. Table 1 shows the results of these tests, where “SACC” stands for “sentence accuracy”. The FAC system achieves a word accuracy of 96.1%, which represents an absolute improvement of 2.6% compared to the “baseline system”. The FAC system reduced the WER by 40% relative. We also tested the FAC system with

the “confusion” extracted from the *TIMIT* database and the confusion given by a phonetician expert (as described in the previous section). These two systems give the same results as the baseline, since the added knowledge (confusion) does not contain any information about the non-native pronunciation variants.

MLLR adaptation was performed using the development database. As described in 3.1, half of the utterances of each speaker were used for the supervised global MLLR adaptation and the other half for the testing.

Table 1: Test results (in %, penalty=-50).

system type	DEL	INS	WACC	SACC
- baseline system	0.4	1.1	93.5	87.2
- fully automated “confusion”	0.2	0.6	96.1	91.1
- <i>TIMIT</i> “confusion”	0.2	0.9	93.9	86.2
- expert “confusion”	0.2	1.2	93.2	85.5
- baseline sys. + MLLR	0.4	1.2	95.0	90.4
- fully automated “confusion” + MLLR	0.1	0.5	97.3	93.5

The good results presented in table 1 could be explained by the nature of the grammar (a strict command grammar). Thus, to be able to fully appreciate the improvement of our approach, we launched tests with a word-loop grammar. Table 2 shows the results of these tests, still with a word-insertion-log-penalty of -50. For the FAC system, improvements are 32.5% relative for the WER and 9.1% absolute for the word accuracy (compared to the baseline results).

The *TIMIT* “confusion” system performed as good as the baseline. On the contrary, the “confusion” given by the phonetician expert improved the WACC by 3% (absolute) and the WER by up to 10% (relative).

Table 2: Test results with a word-loop grammar (in %, penalty=-50).

system type	DEL	INS	WACC	SACC
- baseline system	0.8	8.9	71.1	61.1
- fully automated “confusion”	0.7	5.9	80.2	66.0
- <i>TIMIT</i> “confusion”	1.2	7.9	70.0	57.0
- expert “confusion”	0.5	8.7	74.1	59.5
- baseline sys. + MLLR	0.7	6.1	78.6	69.0
- fully automated “confusion” + MLLR	0.7	4.1	84.5	73.0

4. Discussion and future work

In this paper, we described a fully automated approach to enhance the performance of an existing ASR system with non-native speech. We introduced a novel phonetic confusion concept (associating a phone with a sequence of phones) that deals with English diphthongs. Finally, our method can be combined with speaker adaptation technique. It is possible to perform a MAP or MLLR adaptation for the phone models.

We carried our tests only on non-native English speech uttered by French people. Our next research will be directed on enhancing a unique ASR system performance on heterogeneous non-native English speech. As the European project *HIWIRE* involves Spanish, Italian, Greek and French teams, we will work on non-native English speech uttered by speakers from these origins. We will explore different issues:

- using only the English acoustic models: an English phone would be confused with a sequence of English phones. This approach may be utilized when there are no available acoustic models for each native language.
- taking into account further information while extracting the phonetic “confusion” such as the phonetic or the graphemic contexts: for instance, the pronunciation of a phone may depend on the phones preceding or succeeding it. Furthermore, the grapheme that corresponds to a phone may influence the way a non-native speaker utters it: a non-native speaker may utter the phone corresponding to a grapheme the way it is uttered in its native language.
- developing a meta-ASR system consisting of a set of modified English ASR systems (one for each possible non-native language): the meta-ASR will have a native language detection layer that determines the native language of the speaker. Based on the latter detection, the meta-ASR will use the underlying modified ASR system to perform the speech recognition.

5. Acknowledgments

This work was partially funded by the European project *HIWIRE* (Human Input that Works In Real Environments), contract number 507943, “sixth framework programme, information society technologies”.

6. References

- [1] Stefan Schaden, “Generating non-Native pronunciation lexicons by phonological rule”. In Proc. 15th ICPhs, pp. 2545-2548, Barcelona. 2003.
- [2] K. Livescu and J. Glass, “Lexical modeling of non-native speech for automatic speech recognition”, In Proc. ICASSP, pp 1683-1686, Istanbul, Turkey. 2000.
- [3] S. Goronzy, R. Kompe and S. Rapp, “Generating non-native pronunciation variants for lexicon adaptation”, In Proc. ISCA ITRW Workshop on Adaptation Methods, pp. 143-146, Sophia Antipolis, France. 2001.
- [4] John J. Morgan, “Making a speech recognizer tolerate non-native speech through Gaussian mixture merging”. InSTIL/ICALL 2004 Symposium on Computer Assisted Learning, paper 052, Venice. 2004.
- [5] P. Nguyen, P. Gelin, J.-C. Junqua and J.-T. Chien, “N-best based supervised and unsupervised adaptation for native and non-native speakers in cars”, In Proc. ICASSP, vol. 1, pp. 173-176, Phoenix. March 1999.
- [6] D. Fohr, O. Mella, I. Illina, F. Lauri, C. Cerisara, C. Antoine. “Reconnaissance de la parole pour les locuteurs non natifs en presence de bruit”. In “XXIVmes Journes d’Etude sur la Parole - JEP’02”, pp. 297-300, Nancy, France. 2002.