FULLY AUTOMATED NON-NATIVE SPEECH RECOGNITION USING CONFUSION-BASED ACOUSTIC MODEL INTEGRATION AND GRAPHEMIC CONSTRAINTS

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. We introduce graphemic contraints in the confusion extraction process: the phonetic confusion is established for each couple of 'L2-phone' and the grapheme(s) correspondig to that phone. We claim that prononciation errors may depend on the graphemes related to each phone. 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. The introduction of graphemic contraints in the phonetic confusion allowed further improvements.

1. INTRODUCTION

The drastic drop in performance for automatic speech recognition (ASR) systems when confronted with non-native speech is a well know problem. The main aim of non-native enhancement of ASRs is to make available systems tolerant to pronunciation variants by integrating some extra knowlegde into existing systems (dialects, accents or non-native variants). Approaches differ in the techniques used to extract this knowledge and integrate it into an existing native system.

In [2], studies achieved by human experts on the phonological properties of both spoken language and native language of the speaker allow the extraction of knowledge about non-native speech accent. This knowledge is represented as a set of phone rewriting rules where 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 [3], phonetic confusion is automatically extracted from non-native speech database by aligning the canonical pronunciation of each utterance with its actual pronunciation. This confusion is then used to modify the lexicon by adding all possible phonetic transcriptions of each word dynamically during the recognition phase.

In [5], 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 (piched up 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's phones (for each state of the HMMs). These modified phone models are then used as new models for the spoken language ASR system.

Our new approach is decribed in [1]. We use both native language and non-native language ASRs in order to extract a "one to many" phonetic confusion. This new confusion concept deals with the problem of non-existence of match between some phones of both native and non-native languages. We will describe it briefly in the next sections.

In this paper, we introduce graphemic constraints into the phonetic confusion. Non-native speakers tend to produce phones and to pronounce words (especially unknown words) in the same manner as in their native language. We claim that the pronunciation errors (or variants) a non-native speaker produces depend on graphemes (or the writing of words). For instance, for the English word "*approach*":

- canonical English pronunciation: [ə] [p] [r] [ɔ] [tʃ]
- French speakers pronunciation (significant % of cases): [a] [p] [r] [ɔ] [t] [ʃ]

The upper French pronunciation is given by a French phonetic recognizer. The French phone "a" is phonologically very far from the English phone "ə". French speakers are simply used to pronounce the grapheme (character) "a" as the French phone "a".

Furthermore, pronunciation errors for the same phone depend on the graphemes related to that phone. Thus, we claim that the phonetic confusion could be more accurate if the graphemic constraints (for phones) are taken into account. The same phone (spoken language) may be mis-pronounced in different manners depending on the graphemes corresponding to that phone. To illustrate this, let's consider the English phone "ə".

Table 1. *Pronunciation of English phone ə*

Word	English	French
Approach	[ə] [p] [r] [ɔ] [tʃ]	[a] [p] [r] [ɔ] [t] [ʃ]
Position	[p] [ə] [ʒ] [i] [ʃ] [ə] [n]	[p] [ɔ] [z] [i] [ʃ] [ɔ] [n]

Table 1 shows that French native speakers produce English phone "ə" as the French phone "a" when it corresponds to the character "A" and as the French phone "ɔ" when it correnponds to the character "O".

2. OUR NEW APPROACH

This method was previously described in [1], we will briefly recall it here. As non-native speakers tend to produce 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.

Besides, some phones of the spoken language may not have corresponding phones in the native language. For instance, the consonant '[ð]' does not exist in French. Furthermore, diphtongs like '[tʃ]' do not exist in French. The latter may however be uttered as the two French phones '[t] [ʒ]' or '[t] [ʃ]', as stated by phonetician experts.

Thus, in our new approach, the confusion involves a phone of the spoken language and a phone sequence of the native language. We automatically extract a confusion between spoken language phones and sequences of phones of the native language using both languages' ASRs.

We utilize this confusion by means of HMM modification rather than by lexicon modification. The introduction of confusion knowledge into the lexicon may result in an excessive growth of the lexicon and thus of the search space (see [3]). Furthermore, merging the Gaussian mixture models of each state of the HMM of the confused spoken and native language phones (as in [5]) may deteriorate the coherence of the acoutic 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 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, we carry out a phonetic alignment using the spoken language ASR system and a phonetic recognition using the native language ASR system. 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 has more than its half included in L's time interval. In the example of figure 1, the sequence of native language phones (M_1, M_2) would be associated with phone L.

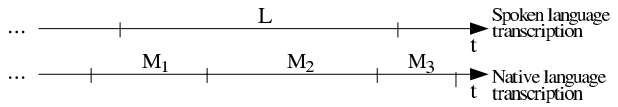

Fig. 1. *Example of time-aligned transcriptions (for the same utterance).*

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):

$$P(L ==> \{M_i\}_{i\in I}) = P((M_i)_{i\in I}|L)$$

$$= \frac{N(L ==> (M_i)_{i\in I})}{N(L)} \quad (1)$$

where $N(L ==> (M_i)_{i\in I})$ is the count of appearances of the underlying association $L ==> (M_i)_{i\in I}$, I a set of indices, and $N(L)$ is 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:

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

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

Here is an example of the confusion rules given by our system for the English diphtong [tʃ] (as in word *church*):

"[tʃ]==>[t] [ʃ]" $P([tʃ]==>[t]\ [ʃ]) = 0.443$
"[tʃ]==>[k] [ʃ]" $P([tʃ]==>[k]\ [ʃ]) = 0.286$
"[tʃ]==>[ʃ]" $P([tʃ]==>[ʃ]) = 0.271$

2.2. HMM integration

In this step, the acoustic models of the phones of the spoken language are modified according to the confusion rules extracted from the previous step. Figure 2 illustrates the HMM structure used in our ASR system for each phone.

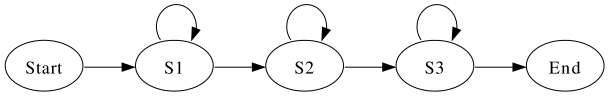

Fig. 2. *Phone HMM model structure.*

For each phone L of the spoken language, a new state path is added to the HMM model of L. These new state paths correspond to the second part of the rules of R'_L (R'_L is the set of selected rules according to the previous section, $R'_L \subseteq R_L$): they are the concatenation of the HMM models of the phones in the right part of the rule.

The transition linking $Start$ state to 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 $Start$ state to each HMM path representing a rule $r \in R'_L$ has a probability of

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

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

3. GRAPHEMIC CONSTRAINTS

As explained above, we consider that the pronunciation errors produced by non-native speakers depend on the writing of pronounced words. Thus, the phonetic confusion should be more accurate if graphemic contarints are taken into account.

The aim is to automatically extract the graphemes linked to the phones for each word of the dictionary. In [2], graphemic constraints and contexts are used. Nevertheless, this phone-grapheme alignment is done manually.

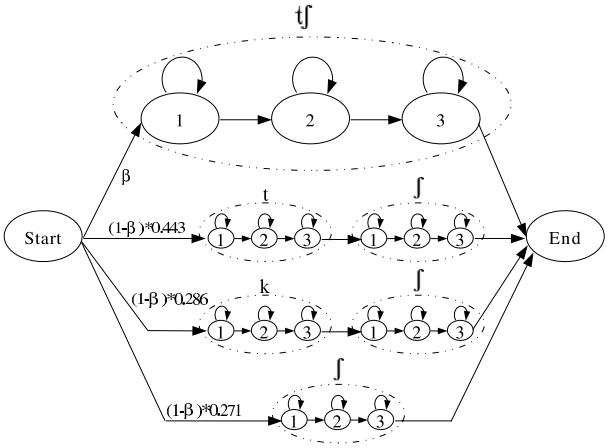

Fig. 3. *Modified HMM model structure for English phone [tʃ].*

3.1. Automatic phone-grapheme alignment

Given the writing of a word and its pronunciation, the task here is to find to which graphemes (characters) each phone corresponds. Even though it seems similar, this task is different from "grapheme to phone". For this latter task, a simple phonetic dictionary may solve the problem for common words. A trained decision tree may be used to cover a larger number of words and unknown names, even random generated words.

In our approach, we use a simple discrete HMM system to perform this alignment on the CMU dictionary[1]. The CMU dictionary was used to train the HMM system. In this discrete HMM system, the characters (graphemes) are the discrete observations and the phones are the HMMs. The HMMs are mono-state discrete HMMs. The trained discrete HMM system can be used to align the graphemes and phones of a word with the classical Baum-Welch algorithm.

3.1.1. HMM system implementation

Here is an illustration of the analogy between our discrete HMM system and classical ASR systems:

- speech data base → training dictionary [2]
- speech utterance → word
- speech observation → discrete symbol [3]
- dictionary → fake dictionary [4]
- grammar → fake word loop grammar
- phones → fake phones [5]
- HMM models → discrete HMM models [6]

As this alignment procedure is meant to be fully automated and applicable to any ASR system, the first step is the analysis of the training dictionary. This allows the extraction of the characters and phones in the dictionary. A translation table between discrete unique symbols and characters is set up. Then, the fake dictionay, the fake grammar and the discrete HMM models are created. These disrete HMM models have a uniform emission probability among all symbols. For each word in the training dictionary, a discrete data file containing the sequence of symbols[7] is created.

The discrete HMM system can then be trained using classical HMM training algorithms. We used *HTK* toolkit and the embedded model reestimation with the tool *HERest*.

A forced alignment is then performed on the whole training dictionary in order to find the associations between phones and graphemes (characters). The second step is to determine the *standard* phone to grapheme associations. This stands for the most often observed phone to grapheme associations for each phone in the training dictionary. A phone to grapheme association a_L (related to phone L) is retained if it satisfies equation 4. This selection avoids erroneous associations resulting from recognition errors or from errors in the training dictionary itself (as this dictionary is hand made).

$$N(a_L) \geq \gamma \, \Sigma_{a'_L \in A_L} N(a'_L) \tag{4}$$

where A_L is the set of phone to grapheme associations for phone L, $N(a_L)$ the count of appearance of the association a_L, and γ a factor.

3.1.2. Applying the graphemic constraints to the ASR system

We propose an approach of the use of the graphemic constraints in the ASR system that is completely transparent to the ASR system itself and the confusion extraction and application method described above. We propose to replace the simple phones of the ASR system by the couples of phones and the graphemes they are related to. This is done by modifying the dictionary of the ASR. The pronunciation of each word is no longer a sequence of simple phones, but it becomes a sequence of phones with their graphemic contraints (characters they are related to in the underlying word). For instance, for the word $used$:

- [i] [u] [z] [d] $\longrightarrow$ [i]-U [u]-U [z]-S [z]-ED

In order to achieve this, a forced alignment is applied to the dictionary of the real ASR system using the trained discrete HMM system described above. It is obvious that the set of characters and phone names used in the dictionary of the real ASR system must be included in those used in the training dictionary.

This way, we obtain phone to grapheme associations for all the phones present in the pronunciation of each word of the dictionary. The phones are renamed (in the pronunciation of each word) as in the example above. If the association obtained for a phone L does not exist in the $standard$ associations (see section 3.1.1), L is kept without graphemic constraint.

The last modification consists in adding HMM models for the newly introduced phones. For each added phone L with a graphemic constraint X, a new HMM model L-X is added to the system. The model for the phone L-X is a copy of the model for the phone L, since, it is obviously the same phone.

3.2. Alignment issues

A single character may be linked to two or more phones. For instance the English word $used$ is pronounced [i] [u] [z] [d]. The use of the straightforward approach described above will lead to the

[1] CMU dictionary version 0.6d
[2] CMU phonetic dictionary
[3] one symbol per character
[4] one word per phone (HMM model)
[5] one phone per HMM model
[6] one per real phone in the training dictionary
[7] that correspond to the characters of the underlying word

unique phone to grapheme association: [i]-U, [u]-S, [z]-E and [d]-D, witch is obviously erroneous. The problem is that for a HMM, observations may not be shared among emitting states. Reverting the states and observations concepts in the discrete system (i.e. considering the characters as the emitting states and the phones as the observations) would not solve the problem as a single phone may be associated with many characters. Rather, we have chosen to duplicate the observations that the system must process. For the same word *used*, the sequence (U, S, E, D) won't be considered, but rather we have the sequence (U, U, U, S, S, S, E, E, E, D, D, D)[8]. This transformation is performed in the hope that the system will make the following associations: ([i]-U, [u]-UU, [z]-SSS, [d]-EEEDDD) or ([i]-UU, [u]-U, [z]-SSS, [d]-EEEDDD). A post-processing transforms the latter into: ([i]-U, [u]-U, [z]-S, [d]-ED).

Table 2. *phone to grapheme associations for phone [u], extracted from the CMU dictionary.*

associated graphemes	appearances count
[u]-U	5659
[u]-OO	1187
[u]-OU	504
[u]-EW	504
[u]-EU	289
[u]-UE	279

4. EXPERIMENTS

The work presented in this paper has been done in the framework of the Europeean project *HIWIRE* which aims at enhancing ASR 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.

4.1. 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. The 40 French monophone models have been trained on the French database *ESTER* which contains 90 hours of broadcast news. The HMM models used have 128 Gaussian mixtures per state and diagonal covariance matrices. The non-native database contains 21 French speakers with 100 utterances for each, 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 314 words, and the grammar is a command language. We also used a "word-loop grammar".

4.2. Development and results

We tested both the baseline and the "fully automated confusion" systems with the grammar presented in 4.1. Table 3 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. No significant improments were obtained by introducing the graphemic constraints along with the phonetic confusion. Nevertheless, graphemic constraints allowed further significant improvements when using a word-loop grammar.

[8]or a multiplication by any integer greater than 2

Table 3. *Test results (in %).*

system type	WACC	SACC
- baseline system	93.5	87.2
- fully automated "confusion"	**96.1**	**91.1**
- fully automated "confusion" + graphemic confusion	**95.9**	**90.8**

Table 4. *Test results with a word-loop grammar (in %).*

system type	WACC	SACC
- baseline system	71.1	61.1
- fully automated "confusion"	80.2	66.0
- fully automated "confusion" + graphemic confusion	**81.6**	**67.16**

Table 2 shows the phone to grapheme associations given by our system for the phone [u]. Here are some examples of the phone to grapheme alignments given by our system:

- *hotel* [h] [ou] [t] [e] [l] → [h]-H [ou]-O [t]-T [e]-E [l]-L
- *mode* [m] [ou] [d] → [m]-M [ou]-O [d]-DE
- *switch* [s] [w] [i] [tʃ] → [s]-S [w]-W [i]-I [tʃ]-TCH

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] G. Bouselmi, D. Fohr, I. Illina, and J.P. Haton, "Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration". In Proc. Eurospeech/Interspeech, Lisboa, September 2005.

[2] Stefan Schaden, "Generating non-Native pronunciation lexicons by phonological rule". In Proc. ICSLP 2004.

[3] K. Livescu and J. Glass, "Lexical modeling of non-native speech for automatic speech recognition", In Proc. ICASSP, 2000.

[4] S. Goronzy, R. Kompe, and S. Rapp, "Generating non-native pronunciation variants for lexicon adaptation". In Proc. Eurospeech 2001.

[5] J. Morgan, "Making a speech recognizer tolerate non-native speech through Gaussian mixture merging". In Proc. InSTIL/ICALL 2004.

[6] 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.

[7] D. Fohr, O. Mella, I. Illina, F. Lauri, C. Cerisara, C. Antoine. "Reconnaissance de la parole pour les locuteurs non natifs en prsence de bruit". In "XXIVmes Journes d'Etude sur la Parole - JEP'02", Nancy, France. 2002.