Preliminary Investigations in Automatic Recognition of English Sentences Uttered by Italian Children

Matteo Gerosa[a,b]
([a])University of Trento
International Graduate School
I-38050 Povo (Trento) - Italy
gerosa@itc.it

Diego Giuliani[b]
([b])ITC-irst, Centro per la
Ricerca Scientifica e Tecnologica
I-38050 Povo (Trento) - Italy
giuliani@itc.it

Abstract

This paper reports on a initial research activity in the area of non-native children's speech recognition that was carried out by exploiting two children databases, one consisting of speech collected from native English children, the other one consisting of English sentences read by Italian learners of English in the same age range of the native speakers. By exploiting the corpus of native speech a baseline speech recognizer was trained for British English. Recognition results achieved for native speakers were compared with those achieved for Italian children uttering the same English texts. Word error rates achieved for Italian children were 100%-600% higher than those achieved for English children of the same age. Adapting baseline acoustic models by using a small amount of non-native speech from a group of Italian learners of English, resulted in a great improvement in recognition performance on non-native speech.

1 Introduction

Recognition of non-native speech is a very difficult task. By using a system trained on native speech, the recognition accuracy achieved for non-native speakers is usually drastically lower than that observed for the native speakers (Wang et al., 2003; Fisher et al., 2003; Matsunaga et al., 2003). The characteristics of non-native speech are, in fact, very different from those of the native speech used for system training. Non-native speech is characterized by pronunciation errors, accented pronunciation, lexical and syntactical errors that mainly depend on the level of proficiency in the target language as well as on the cross-language interference between the mother language of the speaker and the target language. As a consequence, non-native speech presents higher acoustic and linguistic variabilities with respect to native speech(Tomokiyo, 2000).

In view of a speech-driven language learning application, in this work[1] we investigated ASR for young Italian learners of British English. Two databases were exploited for this purpose, one consisting of native speech collected from English children aged 7-12, the other one consisting of English sentences read by Italian learners of English aged 10-11 and with a basic level of proficiency in English.

First, context-dependent HMMs were trained by using data collected from native English speakers. As expected, recognition results were drastically lower for non-native speakers than for native speakers. A small amount of non-native speech, collected from a group of Italian speakers in the same age range of the test speakers and with a similar level of proficiency in English, was then exploited for acoustic models adaptation in order to improve performance for non-native speakers.

The paper is organized as follows. In Section 2 the speech corpora employed are described. Section 3 is concerned with the speech recognition system setup. Speech recognition experiments are described in Section 4. Finally, conclusions about results achieved are reported in Section 5.

2 Speech databases

Two databases of native and non-native English speech, collected from English and Italian children, were exploited. These speech corpora were collected within the EC funded project PF-Star (http://pfstar.itc.it).

The native speech corpus consists of speech collected by Birmingham University from native English speakers aged between 4 and 14. Children read two lists of digit strings and two lists of ten sentences taken from 'SCRIBE', an anglicised version of the phonetically balanced US TIMIT. In addition, each English child read some simple texts prepared for the recordings of Italian learners of English (these texts are described in the following). Two schools, lo-

[1]This work was partially financed by the Autonomous Province of Trento (Italy) under the project PEACH and by the European Commission under the project PF-STAR.

cated in Birmingham and Malvern, respectively, were involved in data collection. Recordings were performed using an head-worn microphone Emkay 3565. Signals were acquired with a sampling frequency of 22050 Hz and a resolution of 16 bits per sample. In this work, only the portion of the corpus uttered by children aged between 7 and 12 were exploited. Furthermore, speech signals were sampled down to 16 kHz.

The non-native speech corpus consists of English words and short sentences uttered by Italian children aged between 10 and 11. Children were foreign learners of English with a basic level of proficiency. All of them were attending the fifth grade of primary school and had been studying English since the first grade.

Data were collected by ITC-irst in two primary schools following two different recording strategies. In the first school, children were allowed to read texts from the computer screen after having listened to a reference pronunciation, by a native adult speaker, one or more times. In the following we will refer to the data collected in this way as "imitated" speech. Each child was asked to read 44 isolated words, 10 phonetically rich sentences and 10 generic sentences. The list of words and the list of phonetically rich sentences were designed in order to ensure coverage of the British English phonemes. In order to ensure variety, 5 lists of 44 words, 5 lists of 10 phonetically rich sentences and 40 lists of 10 generic sentences were prepared and combined to form 40 different prompt text sets. These prompt text sets were also read, one or more times, by the native English speakers as mentioned above.

In the second school no reference pronunciation was available. However, prompt texts were introduced to the children by the teacher some days before the recording sessions. In this case, each child was asked to read a list of 50 isolated words and a list of 25 phonetically rich sentences. Two lists of 50 words and two lists of 25 phonetically rich sentences were prepared and combined to form two prompt text sets. With few exceptions, isolated words and phonetically rich sentences included in these two prompt text sets were taken from the prompt texts used in the first school. We will refer to the data collected in the second school as "read" speech.

In both schools recording were performed using an head-worn microphone Shure SM10A, signals were acquired with a sampling frequency of 16 KHz and a resolution of 16 bits per sample.

Main characteristics of the native and non-native databases are reported in Table 1. The native corpus was partitioned into training, development and evaluation sets, while the non-native corpus was partitioned into adaptation, development and evaluation sets. The development and the evaluation sets for native and non-native databases were designed so that they are formed by repetitions of almost the same English texts.

corpus		Non-Native		Native
language		English		English
training/ adapt.	speakers	30		108
	rec. hours	1h:11m		8h:23m
	age	10-11		7-12
develop.	dev. set	Imit.	Read	Native
	speakers	12	12	12
	age	10	10	10
eval.	eval. set	Imit.	Read	Native
	speakers	12	12	12
	age	10	10	10

Table 1: Statistics of speech corpora used for training and recognition experiments.

3 Speech recognition setup

The parametric representation of speech signals was obtained as follows. Each speech frame was parameterized into 12 mel frequency cepstral coefficients and log-energy. These coefficients plus their first and second order time derivatives were combined to form 39-dimensional observation vectors. Cepstral mean subtraction was performed on an utterance by utterance basis while log-energy was normalized with respect to the maximum value in the utterance.

Recognition experiments were carried out with context-independent (CI), word-internal triphone (WI) and cross-word triphone (CW) HMMs. Output distributions associated with HMMs states were modeled with mixtures of Gaussian densities having diagonal covariance matrices.

For context-independent acoustic modeling, a set of 44 phonetic units, corresponding to the phonemes of British English, were modeled with three-state left-to-right HMMs. Gaussian mixtures associated with the states of a given model shared the same pool of Gaussian densities. During training up to 256 Gaussian densities were allocated for each context-independent HMM depending on the training data available.

For context-dependent word-internal modeling, a phonetically tied mixtures (PTM) scheme was adopted for tying Gaussian mixture components of triphone HMMs. With this tying scheme, all triphones corresponding to a base phoneme have mixtures whose components belong to the same phoneme-dependent pool of Gaussians. Each phoeneme-dependent pool

of Gaussians was allowed to have up to 256 components. In addition back-off models for unseen triphones (i.e. diphones and context-independent models) were constructed following the procedure described in (F. Brugnara, 2001).

For context-dependent cross-word modeling, a Phonetic Decision Tree (PDT) was used for tying the states of triphone HMMs (F. Brugnara, 2001). Output distributions associated with HMMs states were modeled with mixtures with up to 8 Gaussian densities.

In all model sets, "silence" was modeled with a single state HMM. Recognition experiments concerned both isolated word and continuous speech recognition with a close vocabulary. For isolated word recognition a simple finite state network (FSN) grammar with 239 words was used. For continuous speech recognition an unigram language model, implemented with a word-loop FSN grammar and having 636 words, was adopted. Relative frequency for each word was calculated on the training data set. The development set was used to estimate the balancing factor between acoustic and language models.

4 Recognition experiments

4.1 Baseline results

Three different HMM sets were trained by exploiting native speech. This resulted in context-independent, word-internal triphone and cross-word triphone HMM sets having 9150, 9130 and 9815 Gaussian densities, respectively.

Recognition results obtained on native speech with the three HMMs sets are reported on Table 2. The word error rate (WER) is reported separately for the isolated word and continuous speech recognition tasks.

	Isolated Words	Sentences
CI	13.9	36.8
WI	7.4	26.7
CW	7.8	27.3

Table 2: Recognition results in WER (%) on the native evaluation set using context-independent and context-dependent HMMs trained on native speech.

Results show that context-dependent HMMs perform significantly better than the context-independent HMMs. The differences in performance by adopting WI and CW HMMs are relatively small. WI models outperform CW models by 5% WER relative for the isolated word recognition task, while for connected speech recognition the difference is not significant.

Then, recognition experiments were carried out on non-native speech. In Table 3 recognition results are reported separately for the "imitated" and the "read" portion of the non-native evaluation set ("imitated" and "read" refer to the strategy adopted during recording sessions, as described in Section 2).

	Isolated Words		Sentences	
	Imitated	Read	Imitated	Read
CI	44.3	50.3	66.0	69.4
WI	38.8	48.2	57.2	60.0
CW	41.7	54.5	59.7	60.3

Table 3: Recognition results in WER (%) on the non-native evaluation set using context-independent and context-dependent HMMs trained on native speech.

By comparing results reported in Tables 3, we note that WI HMMs outperforms both CW and context-independent HMMs. This is coherent with results reported in the literature on automatic recognition of non-native adults' speech (Morton, 1999). However, WERs achieved are 100%-600% higher than the ones achieved in the case of native speech. Furthermore, we note that in the case of isolated word recognition, performance achieved on "imitated" speech is significantly higher than that achieved on "read" speech for all the HMMs sets considered. In the case of continuous speech recognition, the difference in performance between "read" and "imitated" speech is much smaller. Since the texts uttered in the two portions of the evaluation set are largely the same and the level of proficiency in English among children of the two schools was similar, differences in recognition performance can be mainly attributed to the adopted recording strategy. In fact, a clear difference in pronunciation can be perceived listening to recordings in the "imitated" and "read" portions of the corpus, especially in case of single word utterances. When children uttered after having listened to the reference pronunciation, they were really able to pronounce short sentences following closely the reference pronunciations. However, when uttering longer sentences children paid more attention in reproducing the intonation of the utterances rather than in the correct pronunciation of the single phones especially for less familiar sentences or sentences with unknown words.

4.2 Adaptation experiments

WI HMMs trained with native speech were adapted to better fit with the characteristics of non-native speakers by exploiting the adaptation set, 1h:11m of speech from 30 children of age 10-11, of the non-native speech corpus. Model adaptation was performed by using two

widely used approaches: maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995) and maximum a posteriori (MAP) (Gauvain and Lee, 1994) adaptation. In both cases only the means of the Gaussian components of the HMMs were adapted.

Recognition results achieved on the non-native evaluation set are reported in Table 4.

	Isolated Words		Sentences	
	Imitated	Read	Imitated	Read
MAP	15.9	23.5	34.3	40.5
MLLR	13.1	18.0	31.8	35.4

Table 4: Recognition results in WER (%) on the two portions of the non-native test set performing MAP and MLLR adaptation of WI HMMs trained on native speech.

By comparing these results with those reported in Table 3, a tangible performance improvement can be observed for both MLLR and MAP adaptation. However, MLLR outperforms MAP adaptation. During the adaptation process MLLR exploited a regression class tree for dynamic allocation of regression classes. With the adaptation data available about 400 regression classes were defined and the corresponding transformations estimated and applied to the Gaussian means. In principle, having more adaptation data MAP adaptation should outperform MLLR adaptation.

With adapted WI HMMs, in the case of isolated word recognition, the WER is reduced by 66.2% and 62.6% relative for "imitated" and "read" speech, respectively. For connected speech recognition, the reduction in WER is 44.4% and 42.0% relative for "imitated" and "read" speech, respectively. It can be noted that even if performance improves on both portions of the evaluation set, recognition of "imitated" speech remains still more accurate than recognition of "read" speech.

5 Conclusions

We have investigated automatic recognition of non-native children's speech collected from Italian learners of English aged between 10 and 11 and with a basic knowledge of the English language. Recognition results achieved by using acoustic models trained on speech collected from native children, showed that the word error rates achieved for Italian children were 100%-600% higher than those achieved for English children of the same age. This confirmed that there were systematic differences between native and non-native speech.

Adaptation of word-internal triphones HMMs by exploiting a small amount of non-native speech, collected from children having the same age and a similar level of proficiency in English of the test speakers, showed to greatly improve recognition performance for non-native speakers.

Future work will investigate the use of speaker normalization techniques, in addition to speaker-independent acoustic models adaptation, to better cope with differences between native and non-native speech.

References

F. Brugnara. 2001. Model Agglomeration for Context-Dependent Acoustic Modeling. In *Proc. of EUROSPEECH*, Aalborg, Denmark, Sept.

V. Fisher, E. Janke, and S. Kunzmann. 2003. Recent Progress with the Decoding of Non-Native Speech with Multilingual Acoustic Models. In *Proc. of EUROSPEECH*, pages 3105–3108, Geneva, Switzerland, Sept.

J.-L. Gauvain and C.-H. Lee. 1994. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298.

C. J. Leggetter and P. C. Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185.

A. Matsunaga, A. Ogawa, Y. Yamaguchi, and A. Imamura. 2003. Non-Native English Speech Recognition Using Bilingual English Lexicon and Acoustic Models. In *Proc. of ICASSP*, volume 1, pages 340–343, Hong Kong, April.

R. Morton. 1999. Recognition of Learner Speech. Deliverable D3.3 of the ISLE project, Entropic Cambridge Research Lab. Ltd.

L. M. Tomokiyo. 2000. Handling Non-native Speech in LVCSR: A Preliminary Study. In *Proceedings of the EUROCALL/CALICO/ISCA workshop on Integrating Speech Technology in (Language) Learning (InSTIL)*, Dundee, Scotland, August.

Z. Wang, T. Schultz, and A. Waibel. 2003. Comparison of Acoustic Model Adaptation Techniques on Non-Native Speech. In *Proc. of ICASSP*, volume 1, pages 540–543, Hong Kong, April.