STUDY OF MAXIMUM A POSTERIORI SPEAKER ADAPTATION FOR AUTOMATIC SPEECH RECOGNITION OF PATHOLOGICAL SPEECH

Oscar Saz, Carlos Vaquero, Eduardo Lleida *

Communications Technology Group (GTC)
I3A, University of Zaragoza
{oskarsaz,cvaquero,lleida}@unizar.es

José Manuel Marcos, César Canalís

Colegio Público de Educación
Especial "Alborada", Zaragoza
cpeealborada@gmail.com

ABSTRACT

This paper shows the results achieved by the Maximum A Posteriori (MAP) speaker adaptation method in a task of isolated word Automatic Speech Recognition (ASR) system for 6 young speakers with different types of speech pathologies. In this work, the influence of two important variables in speech recognition and speaker adaptation are studied. On one hand, the performance of the ASR system depending on the acustic unit used by the system (word level unit and context-dependent unit) is analized. On the other hand, how the size of the training set influences the results when using MAP adaptation is studied. The results of this work with the first subset of the Alborada-GTC corpus show that context-dependent units are the most reliable acoustic unit to use, while the use of 2 utterances of the vocabulary recorded in the corpus is enough for getting the best results in speaker adaptation.

1. INTRODUCTION

In the last years, the performance of Automatic Speech Recognition (ASR) systems has improved to very high levels of word accuracy when being used in controlled conditions (noise-free, known acoustic environment and collaborative user). But the performance decays sharply when the system is out of these conditions. This is the case when a user who is suffering a pathology in his/her speech is confronted to an ASR system. Some studies about how the speech is affected when the speaker is suffering any kind of pathology have been made [1]. These studies have pointed mayor variations in the time and frequency domain of the speech as a cause of the failure of the ASR systems when the user suffers any speech pathology. But, patients of any kind of speech pathology are a very likely group of people to use any kind of assistive technology based on speech technologies.

Speaker adaptation has been widely used to introduce ASR systems in our daily life. Several algorithms for speaker adaptation have been developed, Maximum Likelihood (ML) algorithms [2], Maximum A Posteriori (MAP) algorithms [3] and Maximum Likelihood Linear Regression (MLLR) algorithms [4]. MAP adaptation is a good and reliable method for speaker adaptation when sparse data from the speaker is available.

When dealing with speech pathologies, speaker adaptation is the first and basic step to take, but the patients of these pathologies usually find difficult and exhausting to talk in a continous way for a long time. This difficulty to record great amounts of speech for these users forces us to use a quick and reliable algorithm for speaker adaptation like MAP. However, studies on the use of MAP are necessary. The first question to answer is which is the best topology for the acoustic modelling, as it has been studied that the usual topologies may not be the best ones when dealing with pathological speech [5]. The second question to answer is which is the quantity of speech necessary to make an accurate adaptation to the speech of the speaker.

This paper is organized as follows. In Section 2, the Alborada-GTC corpus is introduced. Section 3 is a review of the basis of Maximum A Posteriori (MAP) speaker adaptation. In Section 4 the set of experiments is explained, while the results on MAP adaptation are shown in section 5. Finally, in section 6, the conclusions to this work are given.

2. THE ALBORADA-GTC CORPUS

The acquisition of the Alborada-GTC Corpus is the result of the joint work of the Communications Technology Group of the University of Zaragoza and the teachers and staff of the Public School for Special Education “Alborada”. This work, initial result of further collaborations, was carried away during the first half of year 2006 and finished with the development of the application “Vocaliza”, application for computer-aided speech therapy.

At the end of the process of development of “Vocaliza”, several recordings of children in the school were made, until the completion of the first part of the Alborada-GTC corpus, that it is introduced in this work.

2.1. Speaker composition of the corpus

This first part of the corpus contains recordings from six speakers, whose pathology and situation is explained next:

*This work has been supported by the national project TIN 2005-08660-C04-01

- Speaker BE: Girl, in the range of 10-12 years old, she suffers often and continous dyslalias altogether with sporadic guttural speech. Her speech may be hard to understand for the unused listener.
- Speaker EN: Boy, in the range of 19-21 years old, he suffers slight dyslalias and distorts the pronunciation of some phonemes.
- Speaker ES: Girl, in the range of 10-12 years old, she suffers severe dyslalias and a very weak speech, which does not allow her to properly vocalize. When she is speaking in a continous way, the unused listener might not understand some of the words.
- Speaker JO: Boy, in the range of 16-18 years old, he suffers slight dyslalias, his speech is understable although his speaking rate is very high and utters a high number of disfluencies.
- Speaker RB: Girl, in the range of 13-15 years old, she suffers severe dyslalias and a certain degree of dysphasia. She has problems concerning all levels of language.
- Speaker RG: Girl, in the range of 16-18 years old, she suffers very remarkable dyslalias and she has a non-standard Spanish accent. Her continous speech is hard to understand.

2.2. Phonological composition of the corpus

The set of words chosen for the recordings is the Induced Phonological Register [6]. These 57 words, arranged in terms of their difficulty is a very well-known set of words for the speech therapists in Spain. The set of words contains a rich selection of all the Spanish phonemes in several situations of phonemes boundaries and neighborhood relations.

2.3. Recording parameters

The recordings were made in the classrooms of the "Alborada" School, under the supervision of at least one of the members of the GTC and the surveillance of at least one of the staff of the school. The recordings were made in series of the whole Induce Phonological Register. The tiredness of the children during the process forced to a constant change of the speaker until every one of the 6 speakers recorded 4 series of the Register, which was consider enough speech material for the preliminary work exposed in this paper.

The corpus was recorded with a close-talk wireless microphone (AKG C444L). The use of a close-talk microphone gave us a noise-free recordings, where the average Signal to Noise Ratio is 23.29 dB. The use of a wireless microphone is to avoid that the children were attached to the recording system, so they could feel more comfortable during the recording sessions. Signals were recorded with a 16 kHz sampling frequency and a depth of 16 bits.

3. MAP ADAPTATION

The MAP adaptation [3], though not being the only method available for speaker adaptation, is a well known and reliable method to estimate speaker dependent HMM. This method is based on Bayes theory, and its main feature is the fact that it enables to take advantage of prior information in the training process, thus reducing the data needed to obtain a good speaker dependent acoustic model.

The MAP estimation can be seen as a Bayesian estimation: given a set of n speech feature vectors $X = (x_1, ..., x_n)$, if θ is the parameter vector to be estimated from X, with probability densisty function (pdf) given by $f(X|\theta)$, and g is the prior pdf of θ, it is possible to estimate θ_{MAP} as:

$$\theta_{MAP} = \arg\max_{\theta} g(\theta|X) = \arg\max_{\theta} f(X|\theta)g(\theta) \quad (1)$$

If θ is not known, it can be assumed that $g(\theta)$ gives no information, and (1) then reduces to the Maximum Likelihood (ML) expression.

This estimation method can be applied to estimate multivariate gaussian density functions as well as HMM vector parameters, so it makes possible to estimate speaker dependent or speaker independent acoustic models. To obtain a speaker dependent acoustic model, MAP adaptation usually takes as input a speaker independent model and some utterances from the speaker. The prior information provided by the speaker independent model enables MAP adaptation to obtain good speaker dependent acoustic models using sparse training data. Moreover, even if there is enough training data to use ML estimation, it has been proved that MAP adapted models obtain the same ASR performance in terms of WER as ML adapted models.

4. EXPERIMENTS

For these experiments, a 37 MFCC (Mel Frequency Cepstral Coefficient) parametrization was made, using 12 static parameters, 12 delta parameters, 12 delta-delta parameters and the delta-log-energy. The signals were windowed with a Hamming window of 25 ms. length, with an overlap of 15 ms. Two different acoustic models were used for the baseline, a set of 822 context-dependent units, trained with the whole 16108 utterances of the noise-free training subset of the Spanish SpeechDat-Car [7], and a word model with 57 units, generated by linking the context-dependent units to form whole words. Both models also included two units to model silence and inter-word silence.

Speaker	Word units	Context-dependent units
BE	42.18%	40.79%
EN	34.65%	33.77%
ES	31.14%	31.14%
JO	12.72%	11.40%
RB	58.77%	59.65%
RG	33.33%	32.89%
Average	35.47%	34.94%

Table 1. *Baseline WER results.*

4.1. Baseline per speaker

All the baseline results for the 6 speakers are shown in the Table 1. An average Word Error Rate (WER) of 35% for both models is obtained. Speaker RB clearly exceeds the WER of the rest of speakers, while speaker JO has a much lower WER than the rest.

5. RESULTS ON SPEAKER ADAPTATION

To compute the results of speaker adaptation, a set of 28 different adapted models were estimated for every speaker: 14 word unit based models, and 14 context-dependent unit based models. Each 14 models set was estimated as follows:

- Obtaining 4 different models trained with only one of the series recorded for every speaker. The testing will be made over the other 3 series not used for training.
- Obtaining 6 different models trained with the 6 possible combinations of the 4 series recorded for every speaker, taken 2 by 2. The testing will be made over the other 2 series not used for training.
- Obtaining 4 different models trained with the 4 possible combinations of the 4 series recorded series recorded for every speaker, taken 3 by 3. The testing will be made over the other serie not used for training.

This way, the independence between training and testing data is kept; and the results for the MAP adaptation can be obtained considering training with 1,2 or 3 series as the average of the results of all the models calculated with 1,2 or 3 series.

5.1. Word units

The results obtained considering word unit based models are shown in Table 2. These results are shown for every one of the six speakers of the Alborada-GTC corpus.

The results show that in the best situation, there is an improvement of the 53.82% of reduction of the Word Error Rate. The speaker who gets the highest improvement is BE with a 69.73% reduction of the WER, while the speaker with the poorer performance of the speaker adaptation is JO with a 21.84% reduction of the WER. All the speakers obtain a WER in the gap of 10% to 20%, except the speaker RB, who suffers the most severe pathologies, whose WER is 32.02%.

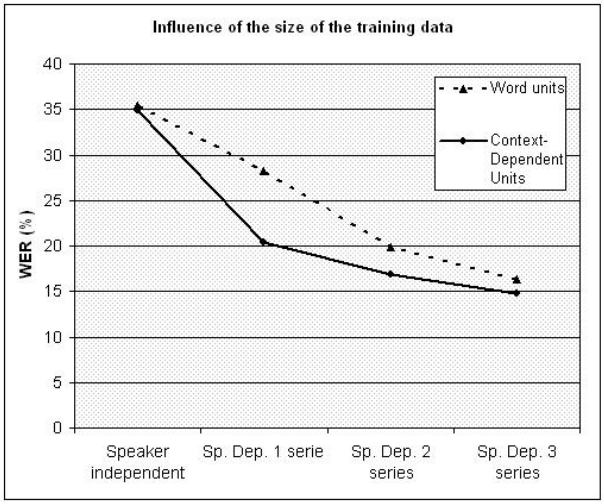

Figure 1. *Influence of the size of the training data.*

5.2. Context-dependent units

The results obtained considering the set of context-dependent units are shown in Table 3.These results are shown for every one of the six speakers of the Alborada-GTC corpus.

The results show that in the best situation, there is an improvement of the 57.73% of reduction of the Word Error Rate, a 4% more reduction than the reduction achived with word units. The speaker who gets the highest improvement is again BE with a 81.72% reduction of the WER, while the speaker with the poorer performance of the speaker adaptation is RB with a 25.73% reduction of the WER. All the speakers obtain a WER lower than 10%, except the speakers RG (12.72%) and RB (44.30%).

Of special interest is the case of the speaker RG, the speaker with the most severe and distinctive pathology all over the set of speakers of the Alborada-GTC corpus. This speaker achieves better results with word models, when the set of training data is big enough (2 or 3 series), than with the context-dependent units: 32.02% of WER with word units versus 44.30% of WER with context-dependent units for the case of the maximum size of training data.

5.3. Influence of the size of the training data

The evaluation of the influence of the size of the training data on the performance of the MAP algorithm for the speaker adaptation is made through the evaluation of the average results over all the speakers for the three possible situations of training (using 1, 2 or 3 series for training). These results are shown on the Figure 1.

The Figure clearly shows that word models are more sensitive to the size of the training data, as the decrease of the WER with this kind of acoustic units is slower than with the context-dependent units. But, with the maximum amount of data available (3 series), the word units obtain only less than 2% of more WER than the context-dependent units (16.38% versus 14.77%).

Speaker	Sp. independent	Sp. Dep. 1 serie	Sp. Dep. 2 series	Sp. Dep. 3 series
BE	42.18%	22.66%	13.60%	12.75%
EN	34.65%	23.69%	19.01%	14.48%
ES	31.14%	27.49%	18.57%	13.16%
JO	12.72%	11.97%	9.94%	10.09%
RB	58.77%	52.05%	36.99%	32.02%
RG	33.33%	31.43%	21.49%	15.79%
Average	35.47%	28.21%	19.93%	16.38%

Table 2. *Speaker adaptation WER results for word units.*

Speaker	Sp. independent	Sp. Dep. 1 serie	Sp. Dep. 2 series	Sp. Dep. 3 series
BE	40.79%	12.58%	9.51%	7.46%
EN	33.77%	17.98%	11.99%	9.65%
ES	31.14%	15.35%	12.57%	8.77%
JO	11.40%	9.79%	7.31%	5.70%
RB	59.65%	47.22%	45.47%	44.30%
RG	32.89%	19.74%	14.48%	12.72%
Average	34.94%	20.44%	16.89%	14.77%

Table 3. *Speaker adaptation WER results for context-dependent units.*

6. CONCLUSIONS

The results exposed in this work show the convenience of using speaker adaptation methods as a way to improve the performance of ASR systems for users with speech pathologies. The results show that MAP adaptation algorithm is as good as to reach a 60% improvement in the reduction of the WER. However, further work should be carried out to make possible a comparison in this task with other speaker adaptation algorithms, like MLLR, that have also shown good performance in several tasks.

As a conclusion to this work about the use of MAP adaptation for pathological speech it is shown that context-dependent units achieve better results than word units. The better performance of the context-dependent units is due to their higher precision to represent the borders between phonemes. A exception to this is the speaker RB, by far the one who is achieving worst results, for whom word units are the best adapted unit. Due to the special characterictics of her speech (dyslalia plus dysphasia), it should be studied if word units are the best units for her kind of speech.

About the size of the training set, good results are obtained when using 2 utterances of every word in the vocabulary (for a total of 114 words), while there is not a much major gain when using 3 utterances of every word. Word units show to be the most sensitive units to the lack of training data, as they achieve much worst results than context-dependent units with only 1 utterance of every word to train; while this difference is reduced when using a greater amount of data for training. This is because the context-dependent units may appear in different words, so they have more data for train with only one serie of the vocabulary of the corpus

7. BIBLIOGRAPHY

[1] Saz O., Miguel A., Lleida E., Ortega A., Buera L., “Study of Time and Frequency Variabilities in Pathological Speech and Error Reduction Methods for Automatic Speech Recognition”, In Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP), Pittsburgh, 2006.

[2] Dempster A. P., Laird N. M., Rubin, D. B., “Maximum likelihood from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society, Series B, vol. 39(1),pp. 1–38, 1977.

[3] Gauvain J.L., Lee C.-H., “ Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains” IEEE Transactions on Speech and Audio Processing, vol.2 (2), pp. 291–298, April 1994.

[4] Legetter C.J., Woodland P.C., “Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden markov models”, Computer Speech and Language, vol. 9, pp. 171–185, 1995.

[5] Deller J.R., Hsu D., Ferrier L.J., “On the use of Hidden Markov Modelling for recognition of dysarthric speech”, Computer Methods and Programs in Biomedicine, 35:125–139, 1991.

[6] Monfort M., Juárez-Sánchez A., “Registro Fonológico Inducido (Trajetas Gráficas)”, Ed. Cepe, Madrid, 1989.

[7] Moreno A., Nogueira A., Sesma A., “Speechdatcar: Spanish” Technical Report SpeechDat.