INTERSPEECH 2009 BRIGHTON

Combination of Acoustic and Lexical Speaker Adaptation for Disordered Speech Recognition

Oscar Saz, Eduardo Lleida, Antonio Miguel

Communications Technology Group (GTC)
Aragón Institute for Engineering Research (I3A)
University of Zaragoza, Zaragoza, Spain
{oskarsaz,lleida,amiguel}@unizar.es

10.21437/Interspeech.2009-196

Abstract

This paper presents an approach to provide of lexical adaptation in Automatic Speech Recognition (ASR) of the disordered speech from a group of young impaired speakers. The outcome of an Acoustic Phonetic Decoder (APD) is used to learn new lexical variants of the 57-word vocabulary and add them to a lexicon personalized to each user. The possibilities of combination of this lexical adaptation with acoustic adaptation achieved through traditional Maximum A Posteriori (MAP) approaches are furtherer explored, and the results show the importance of matching the lexicon in the ASR decoding phase to the lexicon used for the acoustic adaptation.

Index Terms: automatic speech recognition, lexical adaptation, speech disorders

1. Introduction

The variability introduced in their pronunciations by speakers with atypical speech can be so dramatic that canonical transcription of the words in the vocabulary do not match the actual pronunciation of the speaker. This atypical speech could be either the one uttered by a speech impaired individual or by a non-native speaker. As Automatic Speech Recognition (ASR) systems are getting more and more usual in the everyday life (hands-free system, call-centers, etc...), one of their more recent challenges is to provide effective service to this kind of users.

Speaker adaptation is the way in which ASR systems can learn the speech characteristics of a given speaker to improve their performance for that user. However, speaker adaptation usually refers to *acoustic* speaker adaptation; as usually it is understood that the speaker follows the canonical transcription of the words in the vocabulary. As this statement is not completely realistic for all the atypical speakers mentioned above, *lexical* speaker adaptation can be strongly required and has proven to be helpful in many cases.

Strategies for lexical adaptation [1] include the creation of pre-defined rules to model the pronunciation variants of the target speaker or to learn those variants from labeled data from the speaker. While rule-based methods can be good for modeling typical lexical variations in spontaneous speech [2], heavy mispronunciations require data-driven methods to learn these variants [3]. Furthermore, selecting the way in which the ASR system can decode different competitive variants of the same word is also a major issue in lexical adaptation. Acoustic Phonetic Decoding (APD) is a useful tool in terms of lexical adaptation, as it decodes the most likely pronounced sequence of phonemes uttered[4].

This work was supported by national project TIN2008-06856-C05-04 from the Spanish government

Table 1: *Rate of mispronounced phonemes per speaker according to the human labeling*

Speaker	WER	Speaker	WER
Spk01	1.11%	Spk02	21.57%
Spk03	5.22%	Spk04	3.16%
Spk05	43.49%	Spk06	0.68%
Spk07	12.93%	Spk08	30.82%
Spk09	8.22%	Spk10	21.49%
Spk11	6.76%	Spk12	25.69%
Spk13	56.42%	Spk14	8.99%
		Average	17.61%

The present paper aims to achieve lexical adaptation for a group of young speakers with speech impairments due to cognitive disorders. These speakers have been reported to produce a 17.6% of mispronunciations in their speech, resulting in sometimes heavy variations over the canonical pronunciations. The combination of acoustic and lexical adaptation will be the main study of this work, as the interconnection between both can have a significant effect in how to apply them.

This paper is organized as follows: Section 2 will describe the corpus used for this work. In Section 3 the experimental framework and the baseline results will be described. Section 4 will provide the results in acoustic, lexical and combined adaptation for these users, Finally Sections 5 and 6 will show up the discussion and conclusions to this work.

2. Experimental corpus

The corpus used in this work contains speech from 14 young disabled speakers [5] distributed as 7 boys and 7 girls ranging in age from 11 to 21 years old. These speakers suffer from different developmental disorders that affect their language acquisition, resulting in a great number of mispronunciations (substitution and deletions) at the phonetic level. Physiological disorders in their vocal tract components, due to multiple physical impairments, may also affect their production of speech.

The vocabulary recorded from each speaker was the Induced Phonological Register (RFI: Registro Fonológico Inducido) that contains 57 words of special interest for speech therapy [6]. Four sessions of these 57 isolated words were recorded from each speaker, for a total of 3,192 isolated-word utterances, where each word is 5.13 phonemes long in average

All the utterances in the corpus were labeled to detect the mispronunciations made by the speakers. The labeling pro-

Table 2: *Baseline ASR and APD results (Task-dependent models)*

Speaker	WER	PER	Speaker	WER	PER
Spk01	11.40%	41.18%	Spk02	22.81%	42.02%
Spk03	14.91%	45.29%	Spk04	4.39%	29.62%
Spk05	60.09%	68.15%	Spk06	7.46%	33.82%
Spk07	32.02%	49.74%	Spk08	46.49%	54.02%
Spk09	25.00%	41.27%	Spk10	38.16%	53.51%
Spk11	13.60%	34.67%	Spk12	70.61%	76.71%
Spk13	77.19%	67.38%	Spk14	23.25%	38.10%
			Average	31.96%	48.25%

cess followed the next procedure: Three independent labelers (experts in speech technologies or phonetics) were chosen to evaluate a given session from a speaker, marking each phoneme as correct, mispronounced (and, hence, substituted by another phoneme, but without indicating the replacement phoneme) or deleted. The definitive mark for each phoneme was chosen by consensus among the 3 labelers' marks. If necessary, a fourth labeler was required to untie the decision.

The outcome of the labeling in percentage of mispronounced (substituted or deleted) phonemes per speaker is shown on Table 1. The average result is 17.61% of mispronunciations, with six of the speakers above 20% of mispronunciations (Spk13 reaches more than 50%).

To complete this corpus, a group of young unimpaired speakers were also recorded to obtain a parallel corpus of voices representing the speech of individuals in the same range of age (11 to 18 years old) than the impaired speakers. This corpus contains a session of the RFI from 232 young unimpaired speakers, for a total of 13,224 isolated-word utterances.

3. Experimental Framework and Baseline

This Section will introduce the experimental framework and the baseline results in ASR and APD with the proposed corpus. The recognition system was a Hidden Markov Model (HMM)-based Viterbi decoding framework. 25 context-independent acoustic units were trained, each one representing a phoneme of the Spanish language. Each model was a 3-state HMM, where a state was a Gaussian Mixture Model (GMM) with 16 Gaussians. An extra unit to model silence was also used as a 1-state HMM. The feature extraction method was based on the standard ETSI front-end with 39 features per frame (12 cepstral coefficients and the log-energy plus the first and second derivatives).

Baseline acoustic models were trained from adult speech with several adult speech Spanish corpora (Albayzin [7], Spanish SpeeechDat-Car [8] and Domolab [9]) and adapted via Maximum A Posteriori (MAP) adaptation [10] to the unimpaired children speech presented in Section 2. This model is, hence, a task-dependent model, since it is matching the vocabulary, the age range and the acoustic conditions within the target impaired speech corpus.

The APD required a phonotactic language model to learn the way in which phonemes gather to create meaningful syllables and words in the Spanish language. This model was an n-gram model (9,110 3-grams and 628 2-grams) trained from 700,000 sentences of the Spanish subset in the Europarl text corpus [11] that contains transcriptions in different languages of the sessions in the European parliament.

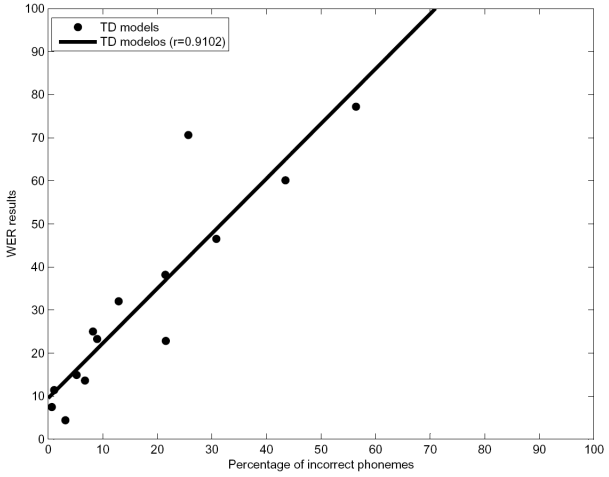

Figure 1: *Correlation of ASR results with speakers mispronunciations'*

Table 3: *Measure of APD for detection of mispronunciations*

Speaker	FRR	FAR	Speaker	FRR	FAR
Spk01	32.73%	23.08%	Spk02	25.76%	9.52%
Spk03	38.48%	22.95%	Spk04	25.91%	16.22%
Spk05	25.91%	12.60%	Spk06	24.91%	25.00%
Spk07	35.99%	13.91%	Spk08	34.16%	8.61%
Spk09	28.82%	11.46%	Spk10	37.30%	12.75%
Spk11	19.56%	17.72%	Spk12	57.49%	14.00%
Spk13	26.92%	11.68%	Spk14	26.15%	11.43%
			Average	31.44%	15.07%

3.1. Baseline Results

The baseline results in ASR and APD for all 14 speakers with the task-dependent acoustic model trained with unimpaired children speech is shown on Table 2, where the first column is presenting the Word Error Rate (WER) of the ASR system and the second column the Phoneme Error Rate (PER) of the APD system. Average result for all speakers was 31.96% WER and 48.25% PER. Results in the same experimental conditions for the unimpaired speakers were 3.31% WER and 15.59% PER, so it could be seen how the disorders suffered by the speakers heavily degraded the performance of the systems

3.2. Correlation with the Speakers' Mispronunciations

As mentioned previously, the WER of the impaired speakers seemed to be heavily influenced by the mispronunciations of the speakers. Assuring this assumption would give further interest to the use of lexical adaptation, because lexical adaptation is especially indicated to avoid the pernicious effect of these mispronunciations. Figure 1 presents the scatter plot of the WER baseline result for each speaker versus the rate of mispronunciations labeled on each speaker by the human experts. Linear regression function is plotted and the regression coefficient (r) is provided. This regression coefficient (0.91) was high enough to indicate the strong correlation in how the speakers' mispronunciations affected the ASR results.

Furthermore, a PER of 48% in APD did not look at a first glance as an accurate predictor of the pronunciation. But this PER was obtained comparing the APD outcome to the canonical transcription of the words (which cannot be considered accurate due to the mispronunciations labeled by the human ex-

Table 4: *WER results with acoustic adaptation with the different approaches*

Speaker	Canonical	APD	Labeling
Spk01	1.75%	12.28%	3.51%
Spk02	10.96%	20.61%	20.18%
Spk03	1.75%	10.09%	3.07%
Spk04	2.19%	4.82%	2.63%
Spk05	47.37%	55.26%	53.95%
Spk06	2.19%	7.02%	2.19%
Spk07	10.96%	28.07%	13.16%
Spk08	31.58%	44.30%	35.09%
Spk09	9.21%	20.61%	13.16%
Spk10	13.16%	39.47%	24.56%
Spk11	3.07%	15.35%	7.02%
Spk12	21.93%	69.30%	28.95%
Spk13	63.16%	73.68%	71.45%
Spk14	7.46%	13.16%	9.21%
Average	16.20%	29.57%	20.58%

perts). To understand how the APD could predict the real pronunciation of the speakers was hence required, and a comparison of the APD results with the human labeling was made in terms of False Rejection Rate (FRR) and False Acceptance Rate (FAR) showed on Table 3. FRR indicated the rate of correctly pronounced phonemes that the APD was deleting or substituting and FAR indicated the rate of mispronounced phonemes that the APD was accepting as the canonical phoneme.

The average rates (31% of FRR and 15% of FAR) indicated that the APD system achieved admissible error prediction rates (comparable to the rates obtained by other pronunciation verification systems with this corpus [12]). Also, the results are well behaved through all speakers and all of them achieved acceptable FRR and FAR rates independently if they produced a higher or lower rate of mispronunciations.

4. Adaptation Experiments

The adaptation experiments were run from 3 different approaches, only acoustic adaptation in subsection 4.1, only lexical adaptation in subsection 4.2 and the combined approach using acoustic and lexical information in subsection 4.3. A thoughtful discussion will be given posteriorly in Section 5.

4.1. Acoustic Adaptation

Acoustic adaptation was performed via MAP adaptation over the task-dependent HMM. A leave-one-out strategy was performed, this is, three sessions were used for re-training and the remaining session was used for testing. Four experiments were, hence, run for each speaker and the final result was the mean of the four WER results.

Three possible transcriptions were fed to the adaptation phase. First one was the canonical transcription of the 57 words in the dictionary, with the results for all speakers given on the first column of Table 4, with an average WER of 16.20%. Second approach used the outcome of the APD seen on Section 3 as the correct transcription, these results are presented on the the second column of Table 4 with an average 29.57% WER. Finally, a "Wizard of Oz" approach was taken using the human labeling of the utterances to discard phonemes from the canonical transcription that were labeled to be mispronounced or deleted. Results are presented on the third column of Table 4 giving an average result of 20.58%.

Table 5: *WER results with lexical adaptation*

Speaker	WER	Speaker	WER
Spk01	10.53%	Spk02	17.98%
Spk03	18.86%	Spk04	5.70%
Spk05	55.70%	Spk06	8.33%
Spk07	24.56%	Spk08	39.91%
Spk09	17.98%	Spk10	27.63%
Spk11	9.21%	Spk12	57.02%
Spk13	64.04%	Spk14	13.6%
		Average	26.50%

Table 6: *WER results with acoustic-lexical adaptation*

Speaker	Canonical	APD	Labeling
Spk01	2.63%	7.46%	2.63%
Spk02	12.72%	15.79%	13.60%
Spk03	3.95%	9.21%	3.95%
Spk04	4.39%	4.82%	4.39%
Spk05	46.05%	51.75%	46.05%
Spk06	2.63%	6.14%	2.63%
Spk07	11.40%	17.54%	12.72%
Spk08	27.63%	36.40%	31.58%
Spk09	9.65%	13.16%	12.72%
Spk10	17.11%	24.56%	15.79%
Spk11	3.51%	8.33%	3.51%
Spk12	30.26%	56.14%	33.77%
Spk13	62.28%	61.84%	59.65%
Spk14	6.58%	13.45%	5.70%
Average	17.20%	23.33%	17.76%

4.2. Lexical Adaptation

The lexical adaptation proposed in this work was based in an APD approach. The outcome of the APD was accepted as the correct phoneme sequence pronounced by the speaker and included in the vocabulary. In the expanded vocabulary, all lexical variants of the same words compete against each other in the Viterbi-based ASR decoding procedure with an equal probability weight for each one. A leave-one-out strategy was taken similarly to the used in the acoustic speaker-dependent system; three sessions from each speaker were decoded with APD, a new lexicon was created for each speaker (with four possible variants: the canonical and the three obtained from the APD) which was used to run ASR with the remaining session. Four experiments were run this way (each one for each evaluation session) and the final result was the mean of the four WER results. Results are provided in Table 5 with an average WER of 26.50%

4.3. Acoustic-Lexical Adaptation

Finally, the possibility of a mixed strategy with acoustic and lexical adaptation was proposed. The outcome of the APD could be fed to the dictionary of the ASR system as proposed in Section 4.2 with any of the three acoustic adapted models proposed in Section 4.1. Using the acoustic model trained with the canonical transcription and the adapted lexicon provided the results on the first column of Table 6 with a 17.20% average on WER; while using the acoustic model trained with the APD-originated transcription and the lexicon adapted with those same transcriptions gives a performance as seen on second column of Table 6 with a 23.33% in average WER. Finally, the adapted models according to the human labeling and the lexical adaptation were used and the results provided in Table 6 with a 17.76% average.

Table 7: *Relative Improvements in WER for different combinations of acoustic and lexical adaptation*

	Acoustic Adaptation			
Lexical	No Adapt	Canonical	APD	Labeling
No Adapt	-	49.31%	7.48 %	35.61%
APD	17.1%	46.18%	27.0%	44.43%

5. Discussion

The discussion of this work was focused on the interconnections between the acoustic and lexical adaptation to improve the performance of the baseline ASR system described on Section 3.

For this discussion, the relative improvements of the WER for the different combinations of acoustic and lexical adaptation are presented in Table 7. This improvement was calculated following Equation 1:

$$Improvement = \frac{WER_{Baseline} - WER_{Adaptation}}{WER_{baseline}} \quad (1)$$

Best results were achieved with the acoustic adaptation using the canonical transcription and no lexical adaptation, although the result adding lexical adaptation was no significantly lower (only 3% less improvement). This points out the bigger effect that the acoustic adaptation produced compared to the lexical adaptation. Improvement with only acoustic adaptation (49%) was significantly higher than the improvement with only lexical adaptation (17%); although the later was remarkable, the acoustic modeling was more powerful to boost the improvement of the ASR system. This could be probably due to the small size of the task, where the acoustic modeling can model in a certain way the different variants of pronunciation.

Lexical adaptation achieved, anyways, a significant reduction in WER in the cases of not using acoustic adaptation (task-dependent acoustic models) and when using acoustic adaptation with the APD-obtained transcriptions (adding a 20% relative improvement) and with the human labeling-based scoring (adding a 9% relative improvement).

These results indicated that lexical adaptation and acoustic adaptation are both useful when they are matching the lexicon used in the recognition phase with the lexicon used in the adaptation phase. When APD or human labels are fed to the transcriptions in the MAP adaptation, the acoustic units only contain data from correctly pronounced segments of speech (or an estimation of them by the APD) and the lexicon in the ASR is matched to recognize these units in new lexical variants of the words.

Hence, this work has proven that lexical adaptation improved the ASR performance without acoustic adaptation or with an unsupervised acoustic adaptation (no knowledge of the prompt to the user is required when using the APD-predicted transcription for acoustic adaptation). This might be useful in many situations in which a supervised acoustic adaptation phase could not be feasible, which is very often the case with heavily impaired individuals, where it is not possible to run long and exhausting speech acquisition sessions for adaptation as required in many applications.

6. Conclusions

As conclusion to this work, a framework for lexical adaptation based on the expansion of the speakers' dictionary with the outcome of an APD system has been tested. The main conclusion of this work has been how lexical adaptation affects when using different frameworks for acoustic adaptation. The need of a degree of matching between the way in which acoustic models are trained and the origin of the new transcriptions for the speakers' vocabulary has been shown. Lexical adaptation when working with a group of impaired speakers as the ones used in this works has proved to achieve major improvements in different situations.

Further work in this area might include a better decision on how to include new lexical variants in the vocabulary. This system could provide some confidence measuring for the phonemes included in the canonical transcription and in the transcription achieved by the APD system to decide more accurately which is the most plausible pronunciation by the speaker. Improving the APD and making it closer to the human labeling would achieve a good performance as proven when the acoustic adaptation is based on those labels.

7. References

[1] H. Strik, "Pronunciation adaptation at the lexical level," in *Proceedings of the ISCA Tutorial & Research Workshop (ITRW) 'Adaptation Methods For Speech Recognition'*, Sophia-Antipolis, France, August 2001, pp. 123–131.

[2] I. Trancoso, D. Caseiro, C. Viana, F. Silva, and I. Mascarenhas, "Pronunciation modeling using Finite State Transducers," in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, September 2003, pp. 1249–1252.

[3] M. Caballero, A. Mariño, A. Nogueiras, "Data driven multidialectal phone set for spanish dialects," in *Proceedings of the 2004 International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, South Korea, August 2004, pp. 837–840.

[4] K.-F. Lee and H.-W. Hon, "Speaker-Independent phone recognition using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[5] O. Saz, W.-R. Rodríguez, E. Lleida, and C. Vaquero, "A novel corpus of children's impaired speech," in *Proceedings of the 2008 Workshop on Children, Computer and Interaction*, Chania, Greece, October 2008.

[6] M. Monfort and A. Juárez-Sánchez, *Registro Fonológico Inducido (Tarjetas Gráficas)*. Madrid, Spain: Ed. Cepe, 1989.

[7] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.-B. M. no, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus," in *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, Berlin, Germany, September 1993, pp. 175–178.

[8] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "Speech Dat Car. A large speech database for automotive environments," in *Proceedings of the II Language Resources European Conference*, Athens, Greece, June 2000.

[9] R. Justo, O. Saz, V. Guijarrubia, A. Miguel, M.-I. Torres, and E. Lleida, "Improving dialogue systems in a home automation environment," in *Proceedings of the First International Conference on Ambient Media and Systems (Ambi-Sys 2008)*, Québec City, Canada, February 2008.

[10] J.-L. Gauvain and C.-H. Lee, "Maximum A Posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[11] P. Koehn, "Europarl: A parallel corpus for statistical Machine Translation," in *Proceedings of the 10th Machine Translation Summit*, Phuket,Thailand, September 2005.

[12] S.-C. Yin, R. Rose, O. Saz, and E. Lleida, "Verifying pronunciation accuracy from speakers with neuromuscular disorders," in *Proceedings of the 10th International Conference on Spoken Language Processing (ICSLP-Interspeech)*, Brisbane, Australia, September 2008, pp. 2218–2221.