Avoiding Speaker Variability in Pronunciation Verification of Children' Disordered Speech

Oscar Saz, Eduardo Lleida, W.-Ricardo Rodríguez
Communications Technology Group (GTC), Aragón Institute for Engineering Research (I3A)
Maria de Luna, 1, University of Zaragoza, Zaragoza, Spain
{oskarsaz,lleida,wricardo}@unizar.es

ABSTRACT

This paper deals with the problematic of speaker variability in a task of pronunciation verification for the speech therapy of children and young adults in Computer-Aided Pronunciation Training (CAPT) tools. The baseline system evaluates two different score normalization techniques: Traditional Test normalization (T-norm), and a novel N-best based normalization that outperforms the first by normalizing to the log-likelihood score of the first alternative phoneme in an unconstrained N-best list. When performing speaker adaptation, the use of all the adaptation data from the speaker improves the performance measured in Equal Error Rate (EER) of these systems compared to the speaker independent systems; but this can be outperformed by more precise models that only adapt to the correctly pronounced phonetic units as labeled by a set of human experts. The best EER obtained in all experiments is 15.63% when using both elements: Score normalization and speaker adaptation. The possibility of automatizing a more precise adaptation without the human intervention is finally proposed and discussed.

Categories and Subject Descriptors

H.5 [**Information Systems**]: Information Interfaces and Presentations; J.2 [**Computer Applications**]: Physical Sciences and Engineering

General Terms

Algorithms, Experimentation

Keywords

Pronunciation Evaluation, Children Speech, Speech Disorders

1. INTRODUCTION

Computer-Aided Pronunciation Training (CAPT) tools aim to improve the pronunciation skills of the user, who might be either a child with speech difficulties, or a foreigner in the process of learning a second language. The technologies underneath these tools are based on the ability of the application to detect phonetic mispronunciations and provide feedback to the user indicating these mistakes and proposing corrective methods for achieving the correct pronunciation [10, 6].

However, the task of pronunciation evaluation for the assessment and correction of phonological mispronunciations in children and young adults [11] faces a major problem in the presence of different sources of variability; being speaker variability one of the more remarkable of them, among others like channel and session variability. This paper aims to evaluate different normalization techniques and speaker adaptation frameworks of acoustic models in a pronunciation verification task, with the final aim of developing CAPT tools, especially oriented to young disabled children. The target language is Spanish, but the proposed systems are language independent and, hence, generalizable to any other language.

This paper is organized as follows: Section 2 will present the pronunciation verification problem, how it resembles the speaker verification problem and which methods can be translated from one to another. Section 3 will present the corpus of disordered speech used for the experiments. Section 4 will provide the experimental framework with the results using all the techniques proposed in score normalization and speaker adaptation. Finally, the discussion and conclusions to this work will be shown on Section 5.

2. THE PRONUNCIATION VERIFICATION PROBLEM

The difficulties arising in pronunciation verification are similar in many facets to the difficulties in traditional speaker verification tasks. In speaker verification, the aim is to decide whether a segment of speech has been uttered by a given speaker or if, on the contrary, an impostor speaker has uttered that segment. In pronunciation verification, the aim is to determine if a certain part of a speech signal corresponds to a given phoneme, or if it has been substituted by the speaker for another phoneme.

There are several sources of variability that mask the speaker features like channel and session variability in the speaker verification process. In pronunciation verification, the phoneme features are masked by sources of variability like speaker and channel variability. Hence, similar techniques used in speaker verification for avoiding the undesired effects of channel variability could be used in pronunciation

ICMI-MLMI'09 Workshop on Child, Computer and Interaction November 5, 2009, Cambridge, MA, USA

Table 1: Speakers in the disordered speech corpus

Speaker	Age	Gender	Correct	Substituted	Deleted	Speaker	Age	Gender	Correct	Substituted	Deleted
*Spk*01	14	Female	98.88%	0.94%	0.17%	*Spk*02	11	Male	78.42%	12.41%	9.16
*Spk*03	21	Male	94.78%	4.54%	0.68%	*Spk*04	21	Female	96.83%	2.05%	1.11%
*Spk*05	18	Male	56.51%	26.11%	17.38%	*Spk*06	17	Male	99.32%	0.51%	0.17%
*Spk*07	18	Male	87.07%	7.36%	5.57%	*Spk*08	19	Male	69.18%	17.72%	13.10%
*Spk*09	11	Female	91.78%	5.31%	2.91%	*Spk*10	15	Female	78.51%	13.10%	8.39%
*Spk*11	20	Female	93.24%	5.15%	2.05%	*Spk*12	18	Male	74.32%	13.96%	11.73%
*Spk*13	13	Female	43.58%	30.48%	25.94%	*Spk*14	11	Female	91.01%	5.14%	3.85%

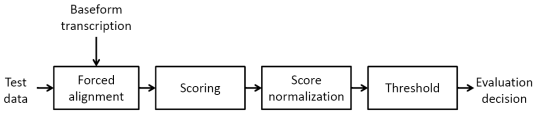

Figure 1: Scoring process

verification for avoiding speaker variability.

The undesired channel variability is erased in speaker verification by a set of proposals like score normalization [1], training of channel adapted models [4] or more novel techniques like Joint Factor Analysis (JFA) [3]. Score normalization and re-training of speaker dependent models are two techniques that can be straightforward applied to the pronunciation verification task, as it was tested during this work.

3. CORPUS

The corpus used in this work contains speech from 14 young disabled speakers with different speech impairments [9]. The distribution of speakers is balanced in terms of age (in the range of 11 years to 21 years old) and gender (7 boys and 7 girls) as seen on Table 1. Speakers have uttered 4 sessions of the words in the Induced Phonological Register (RFI), a very well known set of 57 words designed for speech evaluation in Spanish [5]; giving a total amount of 3,192 isolated word utterances. The sessions were recorded in different days to reflect intra speaker variability.

All these 3,192 utterances were labeled by a group of experts to detect the substitution and deletion of phonemes in these speakers. Three independent labelers evaluated each utterance, and the final label (correct, substituted or deleted) was set by consensus. The rates of correct, substituted and deleted phonemes by speaker are shown in Table 1. The final rate of mispronunciations for all the speakers was 18% of mispronunciations (substituted plus deleted phonemes), what indicated the severe disorders suffered by some of the speakers.

Finally, for the modeling of normal healthy speech in the age of the impaired speakers in the corpus, 232 young speakers without disabilities were recorded uttering one session of the 57 RFI words for a total of 13,224 utterances. The speakers were balanced in the range of age of 10 to 18 years old and in gender.

4. EXPERIMENTS AND RESULTS

The experimental setup was based on the diagram in Figure 1. A Viterbi based forced alignment was made over every input utterance, according to the baseform transcription of

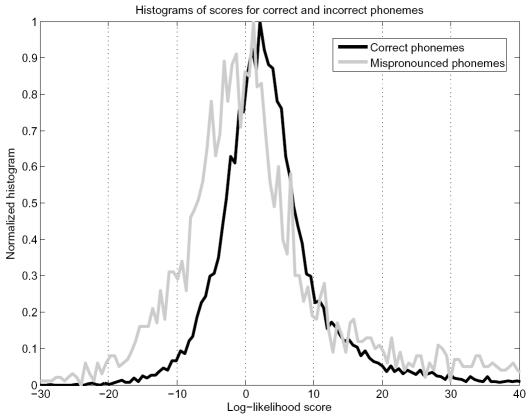

Figure 2: Histogram of the baseline scores

the word to be evaluated. The log-likelihood score was obtained in Equation 1 as the logarithm value of the likelihood probability of the segment of speech s being generated by the model λ averaged by the total number of frames (N_p) assigned to the segment.

$$LL(p) = \frac{log(P(s|\lambda))}{N_p} \tag{1}$$

The likelihood probability of the speech segment and the model was calculated for the N_p frames of the segment as the probability of each frame (t_n with $n = 1...N_p$) in the Gaussian Mixture Model (GMM) ($g = 1...G$) that defined the current state of the phoneme Hidden Markov Model (HMM), as in Equation 2.

$$P(s|\lambda) = \sum_{n=1}^{N_p}(\sum_{g=1}^{G}(p(g)p(t_n|g)) \tag{2}$$

Posteriorly, a score normalization method was applied, and finally, the sigmoid function in Equation 3 reduced the score interval of the log-likelihood score ($LL \Rightarrow [-\infty, +\infty]$) to $LL_{Sigmoid} \Rightarrow [-1, +1]$ and a threshold decided whether each phoneme was evaluated as correct or mispronounced.

$$LL_{Sigmoid}(p) = 2 * (\frac{1}{1 + e^{-LL(p)}}) - 1 \tag{3}$$

The acoustic models used were a set of 27 phoneme Hidden Markov Models (HMM) trained from the unimpaired children speech in the corpus, representing 23 phonemes of Spanish, 2 allophones (glides [j] and [w]), and two silence models. Each model had 3 states whose probability density

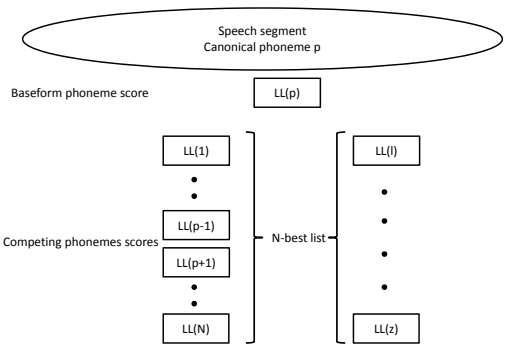

Figure 3: N-best normalization scheme

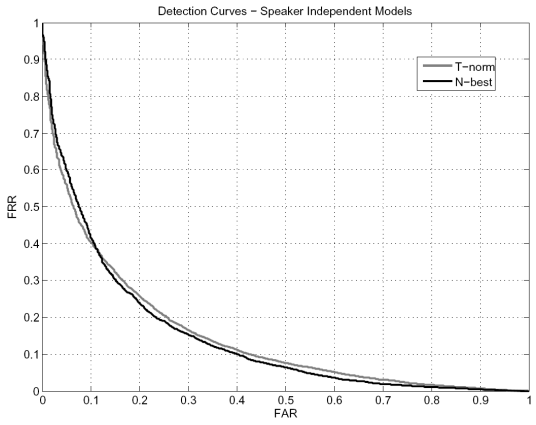

Figure 4: Detection curves with score normalization techniques

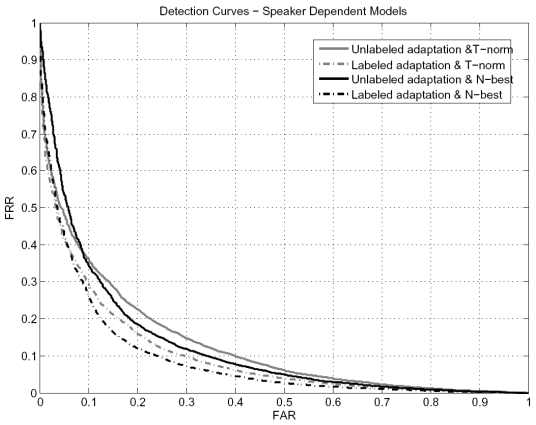

Figure 5: Detection curves with speaker adaptation techniques

function was a Gaussian Mixture Model (GMM) of 32 Gaussians. Every speech input frame was transformed into a 39 Mel Frequency Cepstrum Coefficients (MFCC) containing the first 12 cepstral coefficients and the log-energy of the frame, plus the first and second derivatives of the 13 static parameters.

The log-likelihood scores for the 14 impaired speakers presented a very low separability as it could be appreciated in the histograms of the scores for correct and mispronounced phonemes in Figure 2. This histogram presents the scores prior to the sigmoid function, and it could be seen how both groups (correct vs. mispronounced phonemes) were severely intermingled. A total of 13,472 correct phonemes and 2,880 mispronounced phonemes were evaluated, with their scores calculated, and the Equal Error Rate (EER) for these scores was 43%, close to the worse possible scenario of 50% EER, remarking the dramatic effect of speaker variability.

This preliminary study showed, hence, the need for techniques to avoid the pernicious effects of speaker and channel variability in the pronunciation verification task.

4.1 Score normalization techniques

Two score normalization techniques were proposed as a first way to eliminate speaker variability in the task. The first one was the T-norm approach [1] used in speaker verification. This technique makes a Gaussian normalization of the score achieved by the acoustic model of the target speaker; for this task, the normalization is made over the statistics (mean μ and standard deviation σ) of the scores achieved by the rest of phonemes in the given segment, as in Equation 4

$$LL_{T-norm}(p) = \frac{LL(p) - \mu}{\sigma} \tag{4}$$

A more novel technique was proposed, following the diagram in Figure 3. For a given speech segment, that the forced alignment had assigned to the baseform phoneme p, out of the list of N possible phonemes $(1, ..., N)$, the log-likelihood of phoneme p was calculated as $LL(p)$. The log-likelihoods of all the possible $N-1$ competing phonemes $(1, ..., p-1, p+1, ..., N)$ were also obtained for that segment $(LL(1), ..., LL(p-1), LL(p+1), ..., LL(N))$ and organized in an N-best list according to their values. The log-likelihhod of the first phoneme in the N-best list (for example the phoneme l, with $LL(l)$) was finally subtracted from the log-likelihood of the baseform phoneme ($LL(p)$) for normalization as in Equation 5.

$$LL_{N-best}(p) = LL(p) - LL(l) \tag{5}$$

This method of normalization, in some way resembling cohort normalization [7], was measuring how the baseform phoneme was positioned in terms of the N-best list. Positive values ($LL(p) > LL(l)$) indicated that the baseform phoneme was the most likely for the speech segment (higher normalized score meaning higher difference with the following phoneme in the list); and negative values ($LL(p) < LL(l)$) indicated that one of the competitors was more likely than the baseform phoneme (lower normalized score meaning higher difference between the baseform phoneme and the first phoneme in the N-best list). This score normalization method also obtains automatically an alternative transcription of the prompted word which can be used later for other purposes or for feedback in a language learning tool.

The detection curves for both normalization techniques are provided in Figure 4. The baseline Equal Error Rates (EER) of these proposed systems were 22.80% for T-norm and 21.60% for the novel N-best normalization, with a 5.26% of improvement over the T-norm, indicating the better properties of the proposed score normalization method.

Table 2: EER with speaker dependent models

	Unlabeled adapt.		Labeled adapt.	
	T-norm	N-best	T-norm	N-best
EER	21.29%	19.17%	18.16%	15.63%

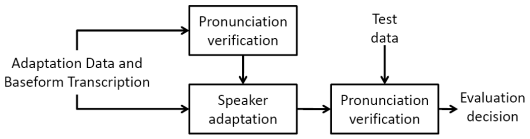

Figure 6: 2-pass adaptation and scoring process

4.2 Speaker adaptation for pronunciation verification

The strategy for speaker adaptation was a set of four leave-one-out experiments where speaker dependent models were trained via the Maximum A Posteriori (MAP) algorithm [2] and the scores of the evaluated sessions were gathered to study the detection curves. MAP algorithm provides a fast and reliable convergence when sufficient data is available. In this case, there are three sessions for adaptation, so every unit appeared a sufficient number of times to provide the good properties of this speaker adaptation. Each model was tested over the remaining session not used for the adaptation.

Two approaches for speaker adaptation were tested. In the first one, all the phonemes in the utterance were used for adaptation, without considering that some phonemes might be mispronounced or not. In the second approach, only the phonemes which were considered as correct by the human labelers were fed to the speaker adaptation. Detection curves for both techniques and both normalizations are plotted on Figure 5 and the EER are in Table 2. EER for them were 19.17% for the "unlabeled" adaptation and 15.63% for the "labeled" adaptation ("labeled" meaning that it used information from the labels given by the experts). Both methods provided relative improvements (11.25% and 27.64%) over the speaker independent models but the adaptation that used the experts' labels gave an extra improvement of more than 15%. Furthermore, in both cases the novel N-best normalization outperformed the T-norm for an up to 14% of improvement.

4.3 2-pass system

Once seen that increasing the precision of the acoustic models increased the performance, it was evaluated the possibility of obtaining automatically an estimation of the mispronounced phonemes, without requiring the manual labeling, with the speaker independent pronunciation verification algorithm. A 2-pass system was designed as seen in Figure 6. Speaker adaptation was run over the adaptation data which was previously evaluated by the speaker independent pronunciation verification algorithm; hence, the input to the adaptation were the speech signals, the baseform transcriptions and the evaluation results to discard mispronounced phonemes. Those speaker dependent models were used in the pronunciation verification of the test data. The same 4 leave-on-out experiments were prepared to assure comparability with the previous results.

This 2-pass system was very dependent on the precise operating point in which the first pronunciation verification system could be configured; for that, three operating points where chosen and studied for this system: The first one is the EER point of the speaker independent verification system, the second one is the point with only 10% of false acceptance (with approximately 40% of false rejections) and the third one is the point with 10% of false rejections (with approximately 40% of false acceptances). Working points situated more in the extremes of the detection curve would lead to the situations of no adaptation as all the data is rejected or adaptation with all the data as everything is accepted (situation similar to the "unlabeled" adaptation).

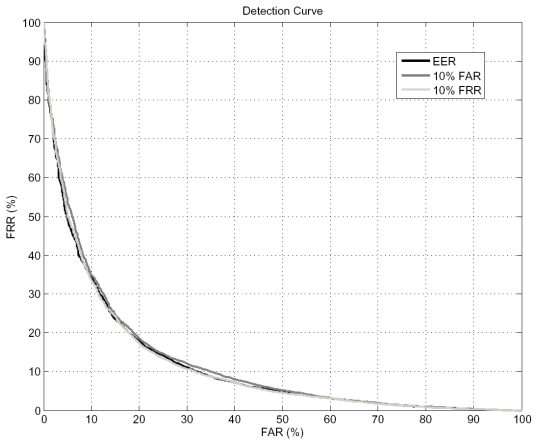

Figure 7: Detection curves for the 2-pass system

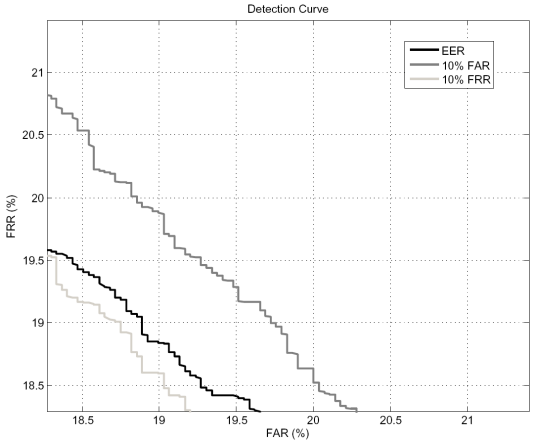

Figure 8: EER zone for the 2-pass system

The three detection curves achieved by the three operating points are plotted in Figure 7 and zoomed around the EER area in Figure 8. The EER for the three working points were 18.91%, 19.38% and 18.82% for the EER, 10% false acceptance and 10% false rejection operating points respectively.

It was seen how the best operation points for the first pronunciation verification phase in the 2-pass system were those who tried to use more data, achieving a 14-15% of improvement over the speaker independent system and a 1.5% of improvement over the "unlabeled" adaptation that used all the data for re-training. This gain had no significance to outperform the system that used all units without considering whether they are correct or not, and it was still far

from the "labeled" adaptation that achieved a 15.63% EER by retraining only with the phonemes considered correct by the human labelers.

5. DISCUSSION AND CONCLUSIONS

Several elements for discussion arose after the experiments run in score normalization techniques and speaker adaptation for pronunciation verification.

Regarding the proposed score normalization techniques, the novel N-best normalization outperformed T-norm with both speaker independent and speaker dependent models. This gain of performance was due to the properties of the phoneme scores in the pronunciation verification task. T-norm hypothesized the Gaussian properties of the different impostor scores and, hence, applied a Gaussian mean and standard deviation normalization.

In phoneme verification, the Gaussian properties of the competing phonemes scores could not be assured. On the contrary, the proposed system aimed to hypothesize an alternative transcription that could be used for providing further information about the pronunciation made by the speaker. However, in further work, the N-best based score normalization technique could be improved by using more information of the N-best list of phonemes instead of only from the first phoneme in the list.

In terms of the speaker adaptation frameworks, the gain of performance achieved by the models adapted only with the correct phonemes (as labeled by the human experts) was due to the better specificity of the trained models, which only had seen correct data and could separate better the phoneme variability in the test utterances. When mispronounced data was fed to the adaptation framework, models lost part of their ability to discriminate correct and mistaken pronunciations, producing worse performance.

Further work in this area might include the improvement of the 2-pass system for speaker adaptation of the phoneme acoustic models with only correctly pronounced data. The results in this work were still far from the better possibility of knowing the mispronounced phonemes by means of the human labeling.

With all this, these results have proven the feasibility of designing CAPT tools based in the proposed systems. An early implementation of them was developed for a Second Language (L2) learning tool, "VocalizaL2", which was tested in a multilingual environment with young students at the Vienna International School (VIS) [8].

6. ACKNOWLEDGEMENTS

This work was supported by national project TIN2008-06856-C05-04.

The authors would like to thank Prof. Richard Rose, Yun Tang and Shou-Chun Yin from McGill University in Montreál (Canada) for their fruitful ideas and discussion for this work.

7. REFERENCES

[1] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1–3):42–54, January 2000.

[2] 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*, 2(2):291–298, 1994.

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Improvements in factor analysis-based speaker verification. In *Proceedings of the 2006 Intl Conf on Acoustics, Speech and Signal Processing (ICASSP)*, pages 113–116, Toulouse, France, May 2006.

[4] P. Kenny, M. Mihoubi, and P. Dumouchel. New MAP estimators for speaker recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, pages 2961–2964, Geneva, Switzerland, September 2003.

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

[6] A. Neri, C. Cucchiarini, and H. Strik. Improving segmental quality in l2 dutch by means of computer assisted pronunciation training with automatic speech recognition. In *Proceedings of CALL 2006*, pages 144–151, Antwerp, Belgium, 2006.

[7] A. Rosenberg, J. Delong, C. Lee, B. Juang, and F. Soong. The use of cohort normalized scores for speaker recognition. In *Proceedings of the 2nd Intl Conference on Spoken Language Processing (ICSLP - Interspeech)*, Banff, Canada, 1992.

[8] O. Saz, V. Rodríguez, E. Lleida, W.-R. Rodríguez, and C. Vaquero. An experience with a Spanish Second Language learning tool in a multilingual environment. In *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education SLaTE*, Wroxall Abbey Estates, United Kingdom, September 2009.

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

[10] J. Tepperman, J. Silva, A. Kazemzadeh1, H. You, S. Lee, A. Alwan, and S. Narayanan. Pronunciation verification of children's speech for automatic literacy assessment. In *Proceedings of the 2006 International Conference on Spoken Language Processing (ICSLP - Interspeech)*, Pittsburgh (PA), USA, September 2006.

[11] C. Vaquero, O. Saz, E. Lleida, and W.-R. Rodríguez. E-inclusion technologies for the speech handicapped. In *Proceedings of the 2008 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4509–4512, Las Vegas (NV), USA, April 2008.