Introducing Difficulty-Levels in Pronunciation Learning

Mark Kane, Joao Cabral, Amalia Zahra and Julie Carson-Berndsen

CNGL, School of Computer Science and Informatics, University College Dublin, Ireland

{mark.kane, amalia.zahra}@ucdconnect.ie {joao.cabral, julie.berndsen}@ucd.ie

Abstract

This paper presents a method to introduce the notion of a student selected difficulty level for the task of pronunciation learning of a second language. Three difficulty levels, *novice, acceptable* and *native*, are constructed using a tri-/bi- and unigram language model each using a specific set of broad phonetic groups determined using underspecification. The evaluation of the experiment is based on a target language of American English where *Nigerian, Hungarian, Indonesian, Syrian, Pakistani* and *Portuguese* students participate. The results show that the addition of difficulty levels returns specific mis-pronunciation information for a student indicating what pronunciation practice is required before they go to the next difficulty level.

Index Terms: pronunciation learning, broad phonetic groups

1. Introduction

The ability of a non-native speaker to improve their second language to that of native status is the proverbial "icing on the cake" that many second language students wish to attain. Coping with learning new words, their correct syntax and past and present tenses is a difficult process for a student or any computer assisted language learning system. While this information is important, the practical issue of pronouncing this information still remains. Pronunciation variation is an essential part of modern language technologies, whether it is accounting for this variation in how a user pronounces an utterance, which could be affected by gender, age, geographical location to the variation that may occur in the system used to detect this variation. Language learning and pronunciation learning tools are not always required in tandem. This separation of learning can be noted in the thespian application of pronunciation learning where the emphasis is on sounding native and not having to know the complete syntax of a language.

Phone-level pronunciation scoring is applied in [1] for language learning of non-native speech that uses a "Goodness of Pronunciation" measure and takes into account individual thresholds for phones; native confidence scores and rejection statistics. This is used in conjunction with a student's speech, force aligned with a pronunciation dictionary of the known practice utterance with respect to a phone recogniser. Explicit error modelling of pronunciation variations is addressed by incorporating correct and common pronunciation errors for each phone.

Pronunciation Learning via Automatic Speech Recognition (PLASER) [2] outlines two speaking exercises: minimal-pair and word exercises. Regarding the latter, a confidence-based score is computed for each phone of a given word and a colour scheme to indicate to the student how well they performed in their pronunciation yielding substantial results.

In [3], broad phonetic groups (BPGs) are used to group similar phones, where phones that share particular characteristics such as manner of articulation belong to the same group. In the same paper it was found that ~75% of misclassified frames were assigned labels within the same group. Phonetic similarity of phones for German and English in [4] shows that there is a notable common confusion between phones, however other factors such as *frequency of occurrence* coupled with their *phonetic neighbourhood* play an important role e.g. the phone */th/* in English does not exist in German.

In [5], the level of pronunciation modelling is attributed to three primary areas of a typical ASR system; *lexicon*, *acoustic models* and *language models*. The approach presented in this paper incorporates BPGs to tackle lexical variants to account for common substitutions in tandem with N-gram modelling where the evaluation of this experiment introduces a 3-tier difficulty level that a student can select. First the student's spoken utterance is force-aligned with the canonical phones constituting the known words in the practice utterance and are then temporally aligned with the recognised phones of the students attempt. If the canonical or recognised phones are equal or belong to the same BPG at the same time, the recognised phone is assumed correct, else they are rejected. This information is then feed back to the student using a phoneme-to-grapheme process to indicate which part of the phrase needs additional practice. This research is part of a web based spoken language coach called MySpeech, a Centre for Next Generation Localisation (CNGL) demonstrator system.

The remainder of the paper is structured as follows: Section 2 introduces pronunciation variation and broad phonetic groups and section 3 details the method in which they are carried out. Section 4 explains the experiment and section 5 discusses the results and future work is highlighted. Finally, conclusions are drawn in section 6.

2. Pronunciation Variation

The primary purpose of this experiment is to note the inter-speaker variation of a non-native participant to a target language by allowing the participant to select a specific difficulty level *novice, acceptable and native* similar to easy, medium and hard. In the decoding process variation is modelled using a data driven and a knowledge based method. Variation is first modelled at a language model level (phone N-gram) and then at a phonetic level by incorporating BPGs, where a difficulty level of *novice* is set by using a tri-gram language model in tandem with several BPGs in comparison to a *native* difficulty level which uses a uni-gram language model in tandem with a sparse amount of BPGs. These BPGs effectively add allowable phone variants instead of introducing specific phone variants for each word in the lexicon. The use of different N-grams is based on the fact that for a participant to achieve a higher score at the *native* setting, the participants acoustic information is required to sound more similar to that of the target language in comparison

to that of the *novice* setting where a tri-gram language model is used.

2.1. Broad Phonetic Groups

Broad Phonetic Groups (BPGs) are constructed by assigning phones that have similar articulatory feature (AF) information to the same group. AFs offer finer grained information of a phone at a specific interval. A phone is considered fully specified for an interval if all of the AFs associated with that phone are present. This approach is based on autosegmental and articulatory phonology underpinned by phonological feature theory [6], [7]. This paradigm was further expanded for speech technology in [8] and [9] and AF-to-phone maps are described in [10] and [11]. Furthermore in [12], AFs which appear simultaneously within the phone, are described as either *on*, *off*, *unspecified* or *unused*. The BPGs presented in this paper are based on underspecification where the grouping of phones depends on partially shared AFs. The BPGs used in tandem with N-gram models are outlined in Table 1.

		novice	acceptable	native
BPGs	phones	tri-gram	bi-gram	uni-gram
#1	n nx	✓	✓	✓
#2	n m	✓		
#3	er axr	✓	✓	✓
#4	th dh t d dx	✓	✓	
#5	p b	✓		
#6	k g	✓		
#7	ch jh	✓		
#8	f v	✓		
#9	s z	✓	✓	✓
#10	sh zh	✓	✓	✓
#11	hh hv	✓	✓	✓
#12	iy ix ih uw ux uh	✓		
#13	ey ow eh ah	✓		
#14	ae ao aa	✓		
#15	iy ih ey eh ae	✓	✓	
#16	ix ux ax axh ah ih	✓	✓	
#17	uw uh ow ah ao aa	✓	✓	
#18	#1 en	✓		
#19	m em	✓		
#20	ng eng	✓		
#21	r #3	✓		
#22	l el	✓		

Table 1: *Broad phonetic groups in tandem with N-gram models indicating difficulty level. 1) alveolar static nasal/flap, 2) bilabial/alveolar nasals 3) syllabic-rhotic 4) alveolar/dental dynamic plosive/fricative/flap 5,6,7,8,9,10,11) voice minimal pair 12) high 13) mid 14) low 15) front 16) centre 17) back 18,19,20,21,22) syllabic pair.*

3. Method

This method consists of three stages to recognise and evaluate the participants spoken attempt of a specific phrase where an overview is shown in Figure 1. The specifics of these stages will now be explained.

3.1. Stage 1

The objective of this stage is to generate the canonical evaluation information for the spoken phrase such as specific word and phone temporal information so that recognised phones of the spoken phrase, also estimated in this stage, can be compared against each other. Evaluation of the participant's pronunciation is based on comparing two phone strings, the canonical and recognised phones and is similar to [1]. As the phrase

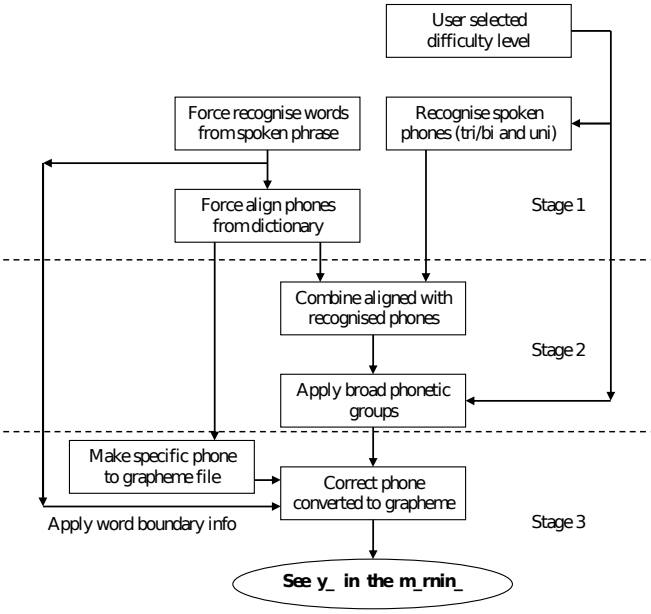

Figure 1: Overview of experiment.

which the participant attempts is selected (in the case of Figure 1; see you in the morning) and therefore known to the system, a specific grammar model is constructed containing the words of the phrase whereby constraining the recognition process to only recognise the sequence of words in the phrase. Once complete, the participant's spoken phrase is force-aligned with a dictionary containing the phones for each word, resulting in a file containing phones and their associated temporal information. This stage also estimates the phones for a participant's spoken utterance using either a tri-/bi- or uni-gram language model depending on the participant's difficulty level selection. The details of the recognition system and corpus used in this experiment are outlined in section 4.1.

3.2. Stage 2

In stage 2, the phones generated from the previous force alignment process are temporally combined with the recognised phones of the participant. This combination process is the same as that shown in Figure 2 similar to [13]. The purpose of temporally combining these phone strings is to find where they are similar at the same time. If similar or belong to the same BPG, the phone is assumed to be correct otherwise it is assumed to be incorrect.

3.3. Stage 3

For each utterance spoken by a participant, a phoneme-to-grapheme representation is created based on the canonical phone sequence of the phrase. The output of stage 2 is then converted to a grapheme sequence where a phone is correctly pronounced. If the phone is incorrectly pronounced, this specific grapheme is highlighted to the participant for further practice.

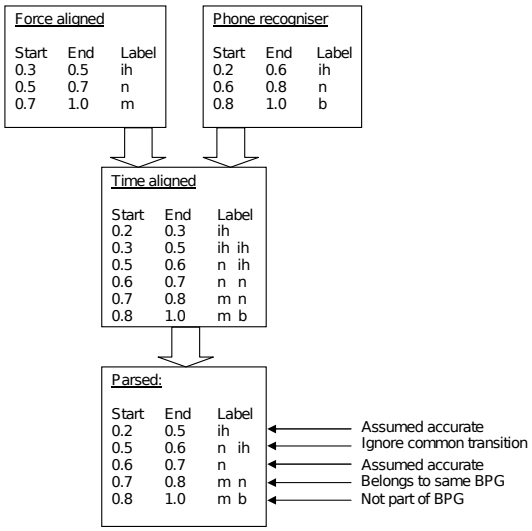

Figure 2: *Example of stage 1 and stage 2 illustrating the combination of force aligned and recognised phones (subset of phones in utterance).*

4. Experiment

In this experiment participants are prompted to repeat an utterance; a common phrase containing several words. Each participant selects a difficulty level and a phrase from a particular category; *time, greetings, food, directions* and *bookings*, in an effort to reproduce the phrase with correct pronunciation. The target language in this experiment is American English whereby 6 participants of different nationalities are evaluated: *Nigerian, Hungarian, Indonesian, Syrian, Pakistani* and *Portuguese*. The repeated utterance is then decoded into a phone string using Hidden Markov models and then compared against the canonical phone string for that phrase based on the aforementioned method.

In total there are 25 phrases where each participant must repeat each phrase three times due to possible intraspeaker variation. It is important to note that the emphasis is not on changing a participant's accent but rather ensuring that each phone is correctly pronounced. Typically a participant receives feedback after each attempt on a phrase, however the format of the experiment presented here is based on examining the participants attempts for each of the 25 phrases, each repeated thrice. As this work is intended to be a web based application, the attempted phrases of all participants were recorded from a laptop in a relatively quiet office where more formal conditions were used for the recording of the phrase prompts. Finally, as there is only one participant for each nationality, each participant's nationality will remain anonymous in Table 2 of Section 5.

4.1. Speech Corpus and Recognition System

The TIMIT speech corpus [14] is used for training (containing 3696 utterances) of the acoustic models and consists of read speech spoken by 630 speakers of American English. All plosives phones represented by a separate closure and associated burst are merged into a single phone e.g. *dcl d* is relabelled as *d* where *dcl* on its own is relabelled as *d*. The HMM based speech recognition system used in this experiment is implemented with HTK [15]. The chosen form of parameterisation of a phone within an utterance is mel frequency cepstral coefficients (MFCCs), with their associated log energy and first and second order regression coefficients. Therefore every frame is represented by 39 coefficients. The MFCCs representing the phones are then used in the calculation of HMM models. The HMMs are context-dependent triphone models that were initially calculated by cloning and re-estimating context-independent monophone models. The triphones states were tied using phonetic feature decision trees for clustering. Each model is comprised of 5 states where only the centre 3 states are emitting. The decoding process is implemented with a tri-/bi- or unigram phone model depending on the difficulty level selected by the participant and is comprised of the prompts for each phrase. Finally, the number of components in each mixture is set to 8 as this was found to be the optimal number for the corpus's test set.

5. Results and Discussion

The results in this section outline the effect of N-gram language models without BPGs and N-gram language models with BPGs as shown in Table 2. This table notes the percent of correctly pronounced phones of a participant across all difficulty levels. From this table it can be noted that participant's six pronunciation of phones has the highest performance across all three difficulty levels where a close second is noted as participants one and three. On *average*, the effect of BPGs at each difficulty level is 4% for *novice*, 16% for *acceptable* and 6.5% for *native*. As shown in Table 1, the *novice* difficulty level has the highest number BPGs, however *novice* BPGs are found to be the least effective due to this level having the highest order N-gram e.g. participant's six correctness without BPGs is 88% and with BPGs is 90%. The *acceptable* BPGs, which number less than half of the previous, contribute the most for phone substitutions e.g. participant's six correctness without BPGs is 66% and with BPGs is 77% whilst the native BPGs were kept to a minimum.

	difficulty level (%)		
participant	novice	acceptable	native
1	(79) 84	(59) 72	(39) 42
2	(78) 83	(53) 63	(35) 39
3	(84) 87	(59) 68	(36) 39
4	(77) 82	(51) 62	(36) 38
5	(86) 87	(54) 66	(36) 38
6	(88) 90	(66) 77	(48) 50

Table 2: *Difficulty level correctness using tri-/bi- and uni-gram. The percentage in parentheses is without BPGs, where the percentage outside parentheses is with BPGs*

participant	difficulty level	# 1	# 2	# 3	# 4	# 5
Portuguese	novice	hh	ae	n	y	w
Portuguese	acceptable	hh	n	uw	m	y
Portuguese	native	ih	ax	ae	d	r
Indonesian	novice	uw	t	n	iy	r
Indonesian	acceptable	t	r	n	m	p
Indonesian	native	t	ax	r	uw	v

Table 3: *Top 5 mis-pronounced phones for Portuguese and Indonesian participants.*

Table 3 shows how the Portuguese participant performs across all difficulty levels for the top 5 mis-pronounced phones (albeit effected by their frequency of occurrence in the phrase prompts). The /hh/ phone is mis-pronounced the most between the *novice* and *acceptable* difficulty levels and is mainly attributed to the fact that this glottal fricative phone does not readily occur in the Portuguese language. As approximant phones /y/ and /w/ are not part of any BPG it is not surprising that they are noted within the top 5 *novice* mis-pronunciations. The main introduction of mis-pronunciations within the *acceptable* level are the nasals as the number of BPGs affecting this manner of articulation is reduced in comparison to the *novice* level. Finally all BPGs relating to vowels are removed at the *native* difficulty level causing this to be the primary source of mis-pronounciations and the area the Portuguese participant needs to improve to sound native after practising at the previous levels.

In the case of the Indonesian participant, there is a consistent trend of mis-pronunciations in the top 5 mis-pronounced phones: /t/, /r/ and /n/ phones. The /r/ phone is easily explained as this phone is typically pronounced as a trill in Indonesia e.g. *cerita* and the /v/ phone is also readily explained as there in no voiced labio-dental in their phone set, only the unvoiced counter part /f/ [16].

5.1. Discussion

It is important to note that the designation of BPGs in this experiment are not a definitive set where the expansion of these groups will be included in future work. Another important aspect of future work is the design of a phonetically balanced prompt set where it is localised to a particular culture using knowledge rich information relating to phones that do and do not exist in the participant's source and target language. This would allow a system, once the source and target language are selected by the participant, to offer prompts that will most benefit the pronunciation learning process.

To further support this pronunciation analysis, pronunciation confidence scores are to be included to account for false-positives and false-negatives. An experiment is to be carried out to compare the confidence scores of force-aligned phones to that of recognised phones and their associated distance where the focus is on implementing a system where a participant can receive a high performance rating when a phrase is pronounced correctly regardless of their accent.

6. Conclusion

This paper introduces the idea of using a difficulty level setting in a pronunciation learning task similar to that of a digital game. A difficulty level setting of either *novice, acceptable* or *native* is selected by a student along with a phrase to attempt. Upon completion of repeating the phrase, specific mis-pronunciations are reported to the student for further practice. This approach allows the student to progress through the learning process at their own pace with the possibility of a competitive aspect introduced when working in groups.

7. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

We would also like to thank John Kane from the Phonetics and Speech Laboratory at Trinity College Dublin for the use of their sound recording facilities for the prompt phrases.

8. References

[1] Witt, S. M. and Young, S. J., 2000. "Phone-level pronunciation scoring and assessment for interactive language learning." Speech Communication, Volume 30, Issues 2-3, Pages 95-108.

[2] Mak, B., Siu, M., Ng, M., Tam, Y. Chan, Y., Chan, K., Leung, K., Ho, S., Chong, F., Wong, J. and Lo, J., 2003. "PLASER: pronunciation learning via automatic speech recognition". HLT-NAACL-EDUC '03 Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing - Volume 2.

[3] Scanlon, P., Ellis, D. and Reilly, R., 2007. "Using broad phonetic group experts for improved speech recognition". IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 3, Pages 803-812.

[4] Kane, M., Mauclair, J. and Carson-Berndsen, J., 2011. Automatic identification of phonetic similarity based on underspecification. In Proceedings of Springer-Verlag Lecture Notes in Artificial Intelligence (LNAI 2011).

[5] Strik, H. and Cucchiarini, C., 1999. "Modeling pronunciation variation for ASR: A survey of the literature." Speech Communication, Volume 29, Pages 225 - 246.

[6] Jakobson R., Fant, G. M. C. and Halle, M., 1952. "Preliminaries to Speech Analysis: The Distinctive Features and their Correlates." MIT Press, Cambridge, MA, U.S.A.

[7] Chomsky, N. and Halle, M., 1968. "The Sound Pattern of English." MIT Press, Cambridge, MA, U.S.A.

[8] Deng, L. and Sun, D.X., 1994. "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features". J. Acoust. Soc. Am. Volume 95, Issue 5, pp. 2702-2719.

[9] Carson-Berndsen, J., 1998. "Time Map Phonology: Finite State Models and Event Logics in Speech Recognition". Kluwer Academic Publishers, Dordrecht.

[10] Scharenborg, O., Wan, V. and Moore, R. K., 2007. "Towards capturing fine phonetic variation in speech using articulatory features." Speech Communication. Volume 49, Issues 10-11, Pages 811-826

[11] Chang, S., Wester, M. and Greenberg, S., 2005. "An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language." Speech Communication, Volume 47, Issue 3, Pages 290-311.

[12] Bates, R. A., Ostendorf, M. and Wright, R. A., 2007. "Symbolic phonetic features for modeling of pronunciation variation." Speech Communication. Volume 49, Issue 2, Pages 83-97 .

[13] Kane, M. and Carson-Berndsen, J., 2011 "Multiple source phoneme recognition aided by articulatory features." In Proceedings of Springer-Verlag Lecture Notes in Computer Science (IEA/AIE)

[14] Garofolo, J., L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, 1993. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM.

[15] Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore and Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland, 2009. "Hidden markov model toolkit (htk)." *http://htk.eng.cam.ac.uk/*, Version 3.4.1.

[16] Zahra, A., Baskoro, S. and Adriani, M., 2009. "Building a pronunciation dictionary for Indonesian speech recognition system." The TCAST workshop, Singapore.