ITERATIVE ENGLISH ACCENT ADAPTATION IN A SPEECH SYNTHESIS SYSTEM

Craig Olinsky (colinsky@mle.ie) *Fred Cummins (fred.Cummins@ucd.ie)*
Media Lab Europe University College Dublin

ABSTRACT

This project documents some initial results of our investigation of the applicability of adaptation techniques and procedures for use in speech synthesis systems – a process whereby a well-trained synthesizer can be transformed to sound "more like" a small target recording set. If we can ensure that the target data set is representative not of an individual speaker, but of a category of accented speech, adaptation of the Synthesis system could be expressed as an iterative procedure where the System learns to modify both its spoken output and linguistic representation to more closely resemble the alternate accent. We summarize the proposed algorithm and present results estimating the rate of pronunciation-learning and generalization possible with this technique, specifically as regards the choice of an appropriately-sized target speech database for training.

1. INTRODUCTION

One of the defining characteristics of concatenative speech systems is the unavoidably close tie between the output voice of the system and that of the source speaker from which the training recordings were made – in speaker quality, voice characteristics, gender, and even spoken genre. There is a secondary relationship between the social and regional characteristics of the source speaker, such as accent and dialect, and the supplementary linguistic knowledge used in the creation of the synthesis framework – phone unit inventories, lexicons, pronunciation rules, etc. The cost and effort in recording, labeling, and processing the large speech databases and supplementary linguistic and lexical data limits the ease with which a variety of voices can be quickly constructed. Inevitably, this leads to a small set of largely “culturally centered”, neutrally-accented voices across the set of languages deemed “commercially viable” for development.

Adaptation in Speech Recognition Systems is a procedure whereby a general-purpose acoustic model trained on a large and varied set of data is transformed to provide better performance on a specific voice through a much smaller target voice training set. Based upon this data, the values of parameters, nodes, weights, or other coefficients representing the acoustic model are shifted “towards” the new information such that the system should exhibit improved performance on data resembling the new training data even though such data was not included in its initial training procedure. Such adaptation is commonly used to personalize commercial recognition systems, transforming speaker-independent systems to improved speaker-dependent systems for individual desktop users. Other uses include the customization of acoustic models for a particular group of users (such as users from the United States of American with a Southern Accent, or non-native Japanese speakers) [13,15]

This project documents some initial results of our investigation of the applicability of adaptation techniques and procedures for use in speech synthesis systems – specifically, a dynamic process whereby a well-trained synthesizer can be transformed to sound "more like" a small target recording set. If we can ensure that the target data set is representative not of an individual speaker, but of a category of accented speech, adaptation of the Synthesis system could be expressed as an iterative process where the System learns to modify both its spoken output *and* linguistic representation to more closely resemble the alternate accent.

For our purposes, we treat accent as the variation across a set of speakers of:

- the ***phonetic inventory*** which comprises the basic building blocks with which things are pronounced;
- a set of ***pronunciation rules*** or examples which dictate how the phonetic units are put together to assign a pronunciation to an orthographic form, and subsequently speak the desired text, and
- a collection of conventionalized ***stress and intonational patterns*** which help provide structure and syntactic / semantic context to the overall produced utterances.

Adaptation in Speech Synthesis System Overview

- *Generate* synthesized utterance from transcript using current synthesizer (letter-to-sound rules, phones, speech database, etc.)
- *Elicit* target recording of the same utterance from a suitable speaker.
- *Compare* target recording to generated source form to determine how the two pronunciations differ.
- *Re-organize* the phone units and speech unit selection process to incorporate differences and info from target recording units.
- *Modify* the lexical entries and letter-to-sound rules, and speech database of the existing synthesizer to produce output that more closely resembles the target utterance.

We have separated the task of synthetic Voice Adaptation into two distinct learning procedures. Broadly speaking, given a fully trained synthetic voice and a set of recorded target utterances, the system first **learns the differences** between its current speech production and that of the target speech, and then **modifies its voice production** based on what it has learned in order to sound more like the target voice. Simultaneous retraining of both phonetic representations and the speech unit database can help avoid the problem of a mismatch between suggested and realized pronunciation caused by introducing a new database of differently accented speech into an existing system.

The primary benefits of dynamic (or even off-line) adaptation of a Speech Synthesis system include the reduction in time, effort, data requirements to build a new voice, and the retained use of declarative and linguistic knowledge (tagging, POS, etc.) already built into an existing system.

In this study, we deal specifically with the representation of pronunciation within the system in terms of the re-assessment (and alteration) of the source phone set for the target data, and similarly the reorganization and modification of the systems pronunciation rules based upon evidence in the target. We have separated this learning process from questions of the actual unit database reorganization and the resultant changes in the output of the concatenative speech system, as the phonetic representation of the desired utterance can be modularized apart from the process of waveform generation. Likely solutions to integrating or simulating the target phonetic units into or from the source unit database come from studies of voice imitation and mimicry [5,17], and voice morphing [8], and will be addressed in a later study.

2. RELATED WORK

Although the goal of voice adaptation follows closely with that of work in recognition systems, our implementation is more directly inspired by efforts in developing multilingual and cross-lingual acoustic models for speech recognizers [2,9,14,16] and targeted model adaptation for non-native accented speech [13,15]. Instead of focusing on speaker variation due to individual differences and voice quality, these studies primarily consider a systematic and generalized difference in phonetic inventory between a set of two languages, employing various means to determine an appropriate mapping between their two phone-sets to determine how acoustic models and trained data can be best shared or borrowed between them. Uebler [14] roughly classifies these methods as ranging from a direct or “na(t)ive” borrowing of all or a subset of an existing phonetic inventory (as one might do when attempting foreign terms from a phrase book), through a “phonetic approach” using external knowledge to map phonemes similar in characteristics such as manner and place of articulation and nasality, to purely data-driven approaches, often employing *confusion matrices* of cross-model acoustic similarity [2]. Predictably [9,14], greater reliance on observed data results in improved accuracy of the mapped model; but the specific language pairings and their overall phonetic similarity provide the greatest variation in results.

Typically these studies have focused exclusively on acoustic mapping, acknowledging, as in [2], to “have assumed that language models, pronunciations, and appropriate acoustic processing are available for the target language, and that only transcribed acoustic training is in short supply.” We find, however, that those languages with limited speech data also pose great difficulty for obtaining quality pronunciation lexicons and rule-sets, and that the logical extension of data-driven phoneme mapping is to additionally allow the system to dynamically target its letter-to-sound rules, in effect “learning” pronunciation rules for the new language, or accent. Most significantly, this addition accounts for the fact that much pronunciation variation across accents and languages isn’t purely a *global* re-mapping of sounds, but instead is highly context-dependent [7].

Automatic learning of letter-to-sound rules has been copiously investigated [1,3,4,12,18], primarily focused on the generalization of such rules from an provided lexicon [1,3] but also addressing augmentation of a recognition dictionary with ambiguous or alternate pronunciation variants to increase recognition [12]. This need is different from synthesis, which necessitates that each word reduce to a single, distinct pronunciation. Contextual rules have long been a favored means of generating pronunciations for text-to-speech systems – they are easily hand-edited, robust to previously unseen words, and provide varying levels of description. This allows us to directly modify the system output at a very high level by modifying or re-ordering rules, without requiring a full batch re-training run. The typical criticism of rule-based approaches – that they “fail miserably” when trained on conflicting examples indicating alternate pronunciations for the same orthographic form [3], we here use to our advantage, for the very point of conflict between predicted and observed data indicates to us exactly which rule must be changed.

In addition to the aforementioned lexical augmentation, phone set redefinition [10,11] provides an additional means of increasing the performance of a recognition system by tuning the acoustic representation to more closely resemble a set of training data. Starting with a seed set of phone units, acoustic data is labeled and re-clustered in an iterative maximum likelihood process, splitting or merging phonetic units when necessary to achieve an optimal balance between the joint likelihood of the training data and the acoustic model.

3. DATA AND METHOD

For our source data, we have trained a concatenative unit-selection synthesizer on a male standard American English speaker based on a labeled recording of the TIMIT speech corpus. This initial system uses a set of pronunciations trained from the CMUDICT using the DARPAbet phone inventory and notation.

Our target data consists of 10 minutes of read speech from three males speakers of Cambridge-Regional British English, selected from from the IViE (English Intonation in the British Isles)[5] speech corpus/database complete with lexical (but not phonetic) transcription. We have used a lexicon derived from the OALD for the purpose of evaluating the results of the pronunciation modification, but it has not played any part in the training procedure.

Following [1,3] we have trained our pronunciation rules into a decision tree from the initial CMUDICT lexicon. This tree implements pronunciation generation as an ordered set of binary decisions based upon orthographic and lexical context, which determine an appropriate phone selection (or alternately multiple or null selections) for each orthographic character.

Converting the lexicon to such a tree allows any changes to pronunciation rules to be reduced to one of the following four procedures:

- Adding to the tree.
- Pruning the tree.
- Modifying a node (or nodes) in the tree.
- Doing nothing.

This process has the advantage of storing the results of multiple iterative changes to our pronunciation rules implicitly within the pronunciation tree itself rather than requiring us to store and process a separate history of all data seen and changes made. As repeated evidence is seen of a pronunciation difference between our source and target speech (that is, as we observe a pronunciation is a generalized form rather than an irregularity or exception) characteristics of this pronunciation form propagate up the tree from a leaf node to an internal branching level.

For each of the utterances in our target set, we generate a corresponding synthesized form using our trained synthesizer, which are used to generate a forced time-alignment and phonetic labeling for the target speech. We then perform what is essentially a low-dimensional *k-means* vector quantization of phonetically labeled, segmented source data followed by a classification of target units within this space. In future studies, this procedure will be replaced with a phone-recognizer more robust to individual speaker recognition; the current approach was chosen specifically to allow reasonable performance given the limits on size and diversity of training data inherent in concatentive synthesis databases.

Comparison of Source and Target Phonetic Composition

```
■ Generate synthesized utterance from transcript using
  current synthesizer (letter-to-sound rules, phones, etc.)
■ Using the synthesized utterance as a guide, segment and
  phonetically label the target utterances into phone units.
■ Plot spectral characteristics of source phone unit database
  in an n-dimensional clustering space, computing centroid(s)
  and deviations across each phonetic label in the current
  inventory.
■ For each observed phone in the target speech:
      Plot its spectral characteristics in the clustering space
      Determine a confidence rating for all possible labels
      If the label of highest confidence is the existing label
          Do nothing
      Else find the phonetic label with the highest score
          If it exceeds a given threshold,
              Re-label the unit as the phone label with best score
          Else if no confidence scores fall within threshold
              Propose a new phone label "x" for this phone
          Proceed to modify pronunciation rules
```

All phone changes with sufficiently high confidence are forwarded on to the rule modification process; which begins by evaluating the orthographic context of the specified phone within the current pronunciation tree. Because the initial pronunciation label assigned by the tree was modified in the preceding comparison procedure, we know that the corrected classification will be different from that predicted by the tree. On this first pass, we change the leaf node of the tree to reflect the predicted phone value (which is either an alternate existing phone unit, or a newly proposed one).

As mentioned, we have set up the rules into a classification tree, ordered downward from most significant to least significant context, ensuring that the each decision node in the tree encodes the *minimum disambiguating context* to separate competing phonetic interpretations. Thus, unless a particular orthographic context is utterly unique (which can be protected against by limiting over-fitting of the rule set during training), each change to a leaf node will impact not one but an entire set of words sharing the relevant context – yet, because we are low in the tree, this context will still be partially constrained.

As phoneme variation across accents is contextual but systematic, there is a strong possibility that the relevant leaf-node changes may cluster in specific areas of the tree, and primarily in the vowel space. In the event that the after they are changed, multiple leaf nodes from a single parent result in an identical phoneme classification, these paths can be pruned and the tree shortened. Pruning, or node removal, can be interpreted as generalization of the pronunciation change. The rate of generalization can be increased, if desired, by storing the predicted information gain for each branch during the creation of the tree, increasing this proportionally above each leaf changed – on the assumption that newer information is more relevant – and pruning lower nodes once the difference in likelihood of following each path reaches a determined threshold.

Node addition, or extension, is opposite of generalization and pruning – rather than changing or supporting an existing rule, it signifies that the change is an exception to the rule. In our implementation, node extension only occurs when the suggested change would reverse or overwrite a recently modified node. Such an occurrence would suggest that the given context – within the new accent – proposes multiple phone unit classifications and thus is not, in fact, minimally descriptive. The context is thus extended by inserting a new decision point sufficient to distinguish the two instances. Although not used here, an external information source such as a tagger of proper names or foreign terms might also be used to put forward a "suggestion" that certain segments be considered as likely rule exceptions, forcing a rule split when they are entered into the tree, whether or not the local node has been recently accessed.

4. RESULTS AND ANALYSIS

Our initial evaluation metrics of the system concern the generalization rate and ability of the rule modification module of the system. This is essential to determining the necessary and sufficient amount of training data to achieve noticeable change in the output pronunciations of the system, as well as in choosing appropriate confidence threshold for initiating changes to the letter-to-sound rules.

The target training databases consists of three speakers reading, consisting of 2995 labeled phones per speaker (or a sum-total of 8985), distributed as follows

52	aa	87	ae	55	ah	43	ao	19	aw	155	ax
52	ay	42	b	6	ch	188	d	97	dh	76	eh
106	er	32	ey	47	f	45	g	63	hh	148	ih
92	iy	13	jh	38	k	147	l	72	m	197	n
18	ng	34	ow	5	oy	46	p	117	r	124	s
13	sh	141	t	7	th	11	uh	42	uw	31	v
76	w	29	y	85	z	266	sil				

Objectively speaking, this is very sparse data. In contrast, the trained pronunciation classification tree has 25,014 rules. If our "phonetic difference" threshold were set so low that every non-silence phoneme in the target speech triggered a unique rule change, this would result in 32.7% of the rules being modified. Even only considering the vowels (the most probable source of accent variation) if every instance in the target set forced a rule change, 12.1%, or about 3000 rules would be modified. Realistically, a change observed in 5-10% of observed vowels would only result in modifications to .7-1.2%, or around 15-300 of the pronunciation rules. Is this enough to significantly impact pronunciations?

Using a wordlist of 12,700 words randomly selected from CMUDICT (10% of the world list), we evaluated the cumulative effect of a maximum of 200 cumulative randomly chosen modifications to the pronunciation rules to determine how large an impact each change had on the resultant pronunciations. We found a fairly consistent average change of 4.975 pronunciations per each change in rules, with the exception of a small number of rules resulting in large-scale change. This suggests that use of a small target training set is reasonable, and that the effect of even a relatively small number of changes will be magnified in the resultant system output. Most significantly, this ratio provides us a guideline by which we can tune the confidence intervals in the comparison module and thus the frequency of modifications to the rule set in order to sustain a learning rate appropriate to any give task or accent pair.

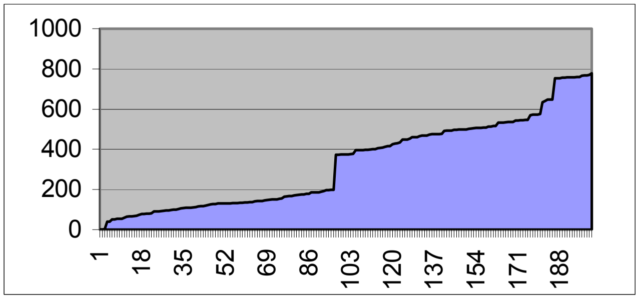

Figure 1: Accumulated Change in Pronunciation Output due to LTS Rule Modification

5. CONCLUSIONS

We have put forward a computationally efficient and real-time capable method for redefining the phonetic output of a pre-trained speech synthesis system, allowing large scale change of phonetic rules with a limited set of training data, while still providing overall gating parameters and mechanisms to prevent over-generalization. Unlike such procedures as phoneme re-mapping, our procedure is data-driven and simultaneously updates both pronunciation rules and phonetic inventory to ensure a proper and accurate relation between the linguistically represented and acoustically generated output of the system.

6. REFERENCES

[1] Black, A.W., K. Lenzo and V. Pagel. "Issues in Building General Letter to Sound Rules." In *Proc. ESCA Workshop on Speech Synthesis* (Australia), pp. 77-80., 1988.

[2] Byrne, W, et al., "Towards Language Independent Acoustic Modeling." *IEEE Workshop on Automatic Speech Recognition and Understanding*, Colorado, December 1999

[3] Daelemans, W. and A. van den Bosch. "Language-Independent Data-Oriented Grapheme-to-Phoneme Conversion," In Van Santen, J., R. Sproat, J. Olive, and J. Hirschberg (eds.) *Progress in Speech Synthesis,* NY: Springer Verlag, 77-90. 1996

[4] Damper, R.I., Y. Marchand, M.J. Adamson, and K. Gustafson, "A Comparison of Letter-To-Sound Conversion Techniques for English Text-To-Speech Synthesis." *Proceedings of the Institute of Acoustics* 20:6, 1998.

[5] Eriksson, A. & P. Wretling. "How flexible is the human voice? - A case study of mimicry." In *Proc. EUROSPEECH '97*, Vol. 2, 1043-1046, 1997.

[6] Grabe, E., Post, B. and Noal, F. "The IViE Corpus." Department of Linguistics, University of Cambridge, 2001. http://www.phon.ox.ac.uk/~esther/ivyweb

[7] Hughes, Arthur and Peter Trudgill. *English Accents and Dialects: An Introduction to Social And Regional Varieties of British English.* London: Edward Arnold Ltd., 1987,

[8] Kain, A, *High Resolution Voice Transformation,* Ph.D. thesis, Oregon Graduate Institute, 2001.

[9] Köhler, J. "Multi-lingual Phoneme Recognition exploiting Acoustic-phonetic Similarities of Sounds" *Proceedings of the 6th International Conference on Speech and Language Processing* (ICSLP), pp. 2195-2198, Philadelphia 1996.

[10] Singh, R., B. Raj and R.M. Stern. "Structured Redefinition of Sound Units by Merging and Splitting for Improved Speech Recognition." *Proceedings of the 6th International Conference on Speech and Language Processing* (ICSLP), Beijing, China, Oct. 16-20 2000.

[11] Singh, R. B.Raj, and R,M. Stern, "Automatic generation of phone-sets and lexical transcriptions", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Istanbul, Turkey, June 5-9, 2000

[12] Sloboda, T and A. Waibel, "Dictionary learning for spontaneous speech recognition." *Proceedings of the International Conference on Spoken Language Processing*, ICSLP 96, Philadelphia, PA 1996

[13] Tomokiyo, L.M. *Recognizing Non-Native Speech: Characterizing and Adapting to Non-Native Usage in Speech Recognition.* Ph.D. thesis, Language Technology Institute, Carnegie Mellon University, 2001.

[14] Uebler, U.. "Speech Recognition in 7 Languages". in *Workshop on Multi-Lingual Interoperability in Speech Technology*, Leusden, 1999.

[15] van Leeuwen, D.A. and R. Orr. "Speech recognition of non-native speech using native and non-native acoustic models." *Proceedings of the MIST workshop*, pp. 23-28, Sept. 1999.

[16] Waibel, A., H. Soltau, T. Schultz, T. Schaaf, F. Metze. "Multilingual Speech Recognition" *Verbmobil: Foundations of Speech-to-Speech Translation,* Wolfgang Wahlster (Ed.), Springer Verlag, 2000.

[17] Wretling, P. & A. Eriksson. "Is articulatory timing speaker specific? - Evidence from imitated voices." In *Proc. FONETIK '98*, 48-52. 1998.

[18] Yvon, François. "Self-learning techniques for grapheme-to-phoneme conversion." *Proceedings of the 2nd Onomastica Research Colloquium*, 1994.