SYMBOLIC SPEAKER ADAPTATION FOR PRONUNCIATION MODELING

Kyung-Tak Lee[1,2], *Lynette Melnar*[1], *Jim Talley*[1]

Motorola Labs, Schaumburg (IL) & Austin (TX), USA[1]
Institut Eurécom, Sophia Antipolis, France[2]
{Kyung-Tak.Lee, Lynette.Melnar, Jim.Talley}@motorola.com

ABSTRACT

This paper presents a method of modeling a speaker's pronunciation of a given language as a blend of "standard" speech and other non-standard speech varieties (regional dialects and foreign accented pronunciation styles) by way of speaker-dependent modification of a lexicon. In this system, a lexicon of Standard American English (SAE) forms, the "canonical" lexicon, is filtered and transformed via a group of speech variety (SV) dependent rule sets into a speaker specific set of pronunciation variants (and associated probabilities) for use during recognition. The relative importance of these rule sets depends on the speaker's pronunciation characteristics and is represented by a Speech Variety Profile (SVP) associated with each speaker. A speaker's individual SVP is acquired through feedback from an adaptation process. Convergence to a speaker's SVP represents adaptation of the lexicon (symbolic adaptation) to those SV-specific forms that speaker is likely to utter.

1. INTRODUCTION

Pronunciation modeling methods are typically applied at the lexicon level and focus on generating new surface form transcriptions to better match pronunciation variability. At the same time, such methods select only the most representative variants in order to limit the risks of lexical confusability.

However, in this general strategy, pronunciation variants reflecting potentially many distinct speech varieties are inevitably omitted, and it would be desirable to increase pronunciation coverage by optimizing the existing pronunciation space. An example of work that addresses this issue is the use of SV-specific pronunciation models ([1]). These methods achieve good pronunciation modeling and limit lexical confusability to only the considered speech variety. Furthermore, they may be combined with existing SV classification methods (*e.g.*, [2]) for multiple pronunciation targeting. However, these methods are designed to activate one single speech variety at a time. It would be better to have more flexibility and the freedom to activate more than one speech variety whenever needed, so that pronunciation characteristics are not represented by a single SV-specific model, but rather a combination of them. This assumption is especially true for speakers who are best characterized by several speech varieties of the same language, but remains valid for any person who speaks with a predominate dialect or accent but sometimes pronounces words in a way better described by some other speech variety. Given this observation, it would, nonetheless, be ill-advised to merge all SV-specific dictionaries since lexical confusability increases with the number of considered speech varieties.

One way to address this issue is to limit the number of pronunciations by weighting the available SV-specific models differently depending on the speaker. The method proposed in this paper simultaneously targets a limited set of speech varieties that are associated with each speaker in his/her Speech Variety Profile (SVP). The SVP is a definition of the speaker's pronunciation characteristics and consists of a list of speaker-associated speech varieties and their corresponding probabilities (cf. section 4.1). The SVP thus functions to dynamically constrain lexicon content to model each individual speaker. The algorithm defining this process is described in the following sections.

2. OVERVIEW

Let us consider a new speaker who wishes to use a speech recognition system for the first time. The system initially has no information about this new user's pronunciation characteristics, but we make the assumption that the speaker is well modeled by a subset of the speech varieties for which the system has existing pronunciation models.

The objective is then to identify the probable speech variety(ies) of the enrolled speaker as accurately as possible. This is done through a *Symbolic Speaker Adaptation* (SSA) process, where the person is asked to utter a set of known sentences. It should be noted that, in contrast to standard Acoustic Speaker Adaptation (ASA), SSA does *not* alter the acoustic models, but rather only modifies the speaker's (very compact) SVP, leaving the acoustic models truly speaker independent.

The adapted profile is then used to expand a baseform, canonical lexicon with new pronunciation variants, the set

of which is constrained by the SV probabilities contained in the speaker's SVP. Each speaker's SVP is saved for future sessions[1].

This method remains an offline process at this stage. Nevertheless, it is quite amenable to online utilization as well - it does not face any more challenge than any other adaptation method that might be used online.

3. PRONUNCIATION MODELS

Two different methods were investigated to expand the canonical (SAE) lexicon with new pronunciation variants. The first method uses generally applicable knowledge-based rules, while the second method uses decision trees derived from the data set of these experiments.

3.1. Rules

A distinct set of rules was defined per speech variety, with each rule tagged with an *a priori* probability of being applied. Selection of rules and probabilities comes from several SV-specific studies in phonetics and phonology as well as reports and pedagogical materials concerned with English-language acquisition by speakers of other languages (*e.g.*, [4]). The following are some examples of SV-specific rules:

N. Inland	/ao/ $\rightarrow$ /aa/ (*e.g.*, "call" $\rightarrow$ /k aa l/)
Indian	/th/ $\rightarrow$ /t/ (*e.g.*, "three" $\rightarrow$ /t r iy/)
British	/aa r#/ $\rightarrow$ /aa#/ (*e.g.*, "car" $\rightarrow$ /k aa/) ("#": word boundary)

3.2. Decision trees

For each speech variety and phone combination a separate tree was trained to predict SV-specific phone(s) from a canonical phone and its left and right contexts. The training method is similar to the one presented in [1]: for each training sentence, the reference words are mapped using the SAE lexicon to a pronunciation string. This phone string is then aligned, using a Dynamic Programming (DP) technique, to an SV-specific transcription of phones selected from a recognition results network. The only difference from [1] is that rather than obtaining the SV-specific transcription candidate set from an unconstrained phone recognizer (which generates too many transcription errors), we first generate a pronunciation network from the baseform transcription(s) using *all* sets of rules mentioned in the previous subsection, and then select the best transcription using Viterbi alignment. Questions used to build the trees concern phonetic features (*e.g.*, front, back, round, ...) for the immediate left and right contexts. The CART algorithm [5] was used to train the decision trees from the DP alignment results.

Generation of pronunciation variants depends on these sets of rules or trees, but the choice of set(s) to be applied and their relative importance is governed by the speaker's SVP. A mathematical formulation of their relationships will be described in section 4.3.

3.3. Limitations

In this initial stage of the project, several constraints were observed that precluded the building of an optimal set of pronunciation models. Perhaps most crucially, only the SAE phone inventory (consisting of 39 symbols) was used to describe pronunciation variability of all speech varieties. It was therefore not possible to account for non-SAE sound distinctions. For example, retroflexion of alveolar consonants is a strong acoustic cue for Indian English, but is not represented in the SAE phone inventory. This limitation prevented both the training of more specific acoustic models and the definition of additional rules to account for these sound differences (decision tree methods were also affected since their training was derived from rule productions).

Next, the *a priori* probabilities assigned to rules derive from reports of general usage in the targeted SV communities and were not re-estimated from the actual data used. Since these values help to guess the probable speech variety(ies) of the enrolled speaker (as will be shown in section 4.2), (likely) inaccuracies in their estimation would have (negative) repercussions throughout the system.

Finally, an overwhelming majority of sentences used to train the decision trees was uttered in SAE or in a phonologically similar SV. Although we consider the remaining sentences to still be sufficient for training the other speech varieties, additional non-SAE data would have been preferable in order to build more reliable pronunciation models.

4. SYMBOLIC SPEAKER ADAPTATION

4.1. Overview

The adaptation process is depicted in Figure 1. The following steps are applied for each enrolled speaker and his/her adaptation sentences:

1. Each word in the adaptation sentence is mapped to its baseform transcription(s) (canonical pronunciation(s)).
2. SV-specific transcriptions are derived from the baseform(s) using all rule sets or decision trees, and used to generate a pronunciation network. For each SV-specific form, a list of symbol transformations is kept.
3. A Viterbi alignment is performed using the network to return the most likely sequence of phones actually uttered by the speaker.

[1] A discussion of automatic speaker identification goes beyond the scope of this paper and is not addressed here, but current existing methods in this field could be suitable for this purpose (see for example [3]).

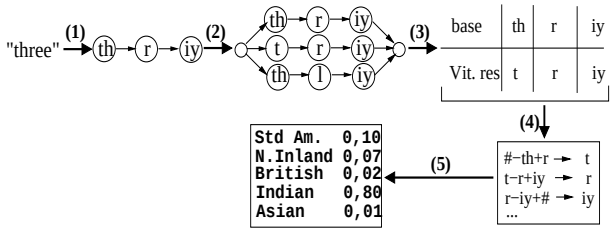

Fig. 1. Symbolic Speaker Adaptation

4. The symbol transformations corresponding to the selected phone sequence are added to a list.
5. Once all adaptation sentences are processed, probabilities for the speaker profile are computed using all transformations found in the list.

An example of a profile after adaptation for an Indian English speaker might be something like this:

Standard American English	0.10
Northern Inland English	0.07
British English	0.02
Indian English	0.80
Asian-accented English	0.01

The next section will describe how the SVP probabilities are computed.

4.2. SVP adaptation

The goal of SSA's process is to calculate the probabilities that a speaker's pronunciation characteristics match each of the speech varieties known by the recognition system. So given a speaker U who uttered some adaptation sentences $\{\sigma\}$, we compute $P(V_i|U,\{\sigma\})$ for all speech varieties V_i. On the assumption that the adaptation sentences contain sufficient information for determining a speaker's speech variety(ies), these probabilities were approximated by the sum of the contributions of all words W_j that the speaker uttered during the adaptation process:

$$P(V_i|U,\{\sigma\}) \approx \sum_{j=1}^{N(W_j)} P(V_i|W_j) \cdot P(W_j) \qquad (1)$$

where $N(W_j)$ is the number of distinct words uttered. In recognition of the fact that lexical words may have multiple canonical (SAE) pronunciations, $P(V_i|W_j)$ is expressed in terms of its canonical pronunciations, or baseforms, B_m:

$$P(V_i|W_j) = \sum_{m=1}^{N(B_m)} P(V_i|B_m) \cdot P(B_m|W_j) \qquad (2)$$

where $N(B_m)$ is the number of baseforms for the word W_j. Let us further develop the term $P(V_i|B_m)$ to model pronunciation variations at the canonical, or baseform, level. By using Bayes' rule and simplifying the problem with the assumption that the phones of a baseform are independent, we have:

$$P(V_i|B_m) = \frac{[\prod_{b=1}^{N(p_b)} P(p_b|V_i)] \cdot P(Vi)}{P(B_m)} \qquad (3)$$

where p_b is a phone (in its left and right contexts) and $N(p_b)$ is the number of phones in the baseform B_m. Each phone in the baseform may be realized as: 1) itself, 2) a different phone (substitution), 3) a sequence of phones (insertion) or 4) a null phone (deletion). By summing over all possible realizations of the baseform phone p_b, we obtain:

$$P(p_b|V_i) = \frac{\sum_s P(V_i|p_b,p_s) \cdot P(p_s|p_b) \cdot P(p_b)}{P(V_i)} \qquad (4)$$

where p_s represents any phone or sequence of phones realized from the baseform phone p_b. After substituting the expression (4) into (3) and some simplifications, we obtain:

$$P(V_i|B_m) = \frac{\prod_{b=1}^{N(p_b)} \sum_s P(V_i|p_b,p_s) \cdot P(p_s|p_b)}{P(V_i)^{N(p_b)-1}} \qquad (5)$$

$P(p_s|p_b)$ is the speaker-dependent probability that measures how often the speaker realizes a phone p_b as p_s; it is obtained by counting the number of times this transformation occurs over all realizations of p_b during the adaptation process: $P(p_s|p_b) = \frac{N(p_s|p_b)}{N(p_b)}$. The first term of the sum, $P(V_i|p_b,p_s)$, is the SV-dependent probability and measures how accurately the same phone transformation $p_b \rightarrow p_s$ targets the speech variety V_i. Using the property of independence between p_b and V_i, and assuming that the speech varieties V_i are disjoint, it can be shown that:

$$P(V_i|p_b,p_s) = \frac{P(p_s|p_b,V_i) \cdot P(V_i)}{\sum_i^{N(V_i)} P(p_s|p_b,V_i) \cdot P(V_i)} \qquad (6)$$

where $N(V_i)$ is the number of speech varieties known by the system. $P(p_s|p_b,V_i)$ is given by either the *a priori* probability of the corresponding rule in the V_i rule set of being applied (if it exists, otherwise the probability equals 0) or the estimation given by the decision tree associated with the SV V_i for the $p_b \rightarrow p_s$ realization.

Finally, adaptation of speaker profiles is given by evaluating the expression seen in (1) with the appropriate substitutions, for each speaker and each speech variety.

4.3. SV-specific form probabilities

Any pronunciation variant derived from a lexical baseform (SAE pronunciation) is assigned a probability of occurrence

to be used during recognition. This subsection describes how they are obtained.

Let us consider a word W phonologically transcribed by $N(B_m)$ baseform pronunciations in the lexicon. We would like to calculate the probability of occurrence of an SV-specific pronunciation S_n given that word, $P(S_n|W)$. Since a word is entirely represented by its baseforms, we can write:

$$P(S_n|W) = \sum_{m=1}^{N(B_m)} P(S_n|B_m) \cdot P(B_m|W) \qquad (7)$$

Pronunciation characteristics of a speaker U (who utters the word W) are represented by $N(V_i)$ speech varieties found in his/her SVP. Since several SVs may accept S_n as a possible output form derived from a baseform B_m, the probability $P(S_n|B_m)$ seen above must take all $N(V_i)$ considered speech varieties V_i into account:

$$P(S_n|B_m) = \sum_{i=1}^{N(V_i)} P(S_n|B_m, V_i) \cdot P'(V_i) \qquad (8)$$

where $P'(V_i) = P(V_i|U, \{\sigma\})$ is the probability that the speech of the speaker U conforms to the i-th speech variety (see section 4.2).

The process to evaluate $P(S_n|B_m, V_i)$ differs between the decision tree and rule methods. The processes are explained respectively in 4.3.1 and 4.3.2.

4.3.1. SV-specific form probabilities using trees

Evaluation of $P(S_n|B_m, V_i)$ using decision trees is quite straightforward. Each phone p_b of the baseform B_m is realized as p_s that represents the same phone p_b, a distinct phone (substitution, deletion) or a group of phones (insertion)[1]. The SV-specific form S_n is obtained by simply concatenating the successive p_s realized from each baseform phone p_b. Assuming that the p_b's are independent, we can write:

$$P(S_n|B_m, V_i) = \prod_{b=1}^{N(p_b)} P(p_s|p_b, V_i) \qquad (9)$$

where $N(p_b)$ is the number of phones in the baseform B_m. Each term of the product is estimated by the decision tree associated with the speech variety V_i and baseform phone p_b.

[1] At this point no special control on the number of phones inserted is necessary, because the SV-specific forms used to train the decision trees are generated by rules (see section 3.2), in practice, that limited the number of phone insertions to one.

4.3.2. SV-specific form probabilities using rules

Let us focus on the rules responsible for the transformation of a baseform B_m to an output form sequence S_n, to see how they influence the probability $P(S_n|B_m, V_i)$. In this framework, each speech variety V_i is associated with a vector (ordered set) of rules $r_i = (r_i^1, r_i^2, ...)$. Each rule r_i^j of the set is *eligible* to transform a sequence of phones only if the sequence matches the *pre-conditions* of the rule, that is, if the sequence contains the focused phone along with any neighbor context(s) required by the rule. Additionally, all rules are considered *optional*, which means that even when a rule is eligible, it is not necessarily applied. To represent these possible rule states in a more compact form, we define a variable $q_i = (q_i^1, q_i^2, ...)$, where each q_i^j represents the state of a rule r_i^j, with three possible values:

1. '0': the rule is not eligible
2. '+': the rule is eligible and applied
3. '-': the rule is eligible, but not applied

We come back now to the process of transformation of a baseform B_m to a SV-specific form S_n, but this time bringing the rules and rule states to the fore. To find the probability $P(S_n|B_m, V_i)$ of equation 8, we are looking for all combinations of rules that successively transform the baseform B_m into the SV-specific form S_n: $B_m \rightarrow X_1 \rightarrow X_2 \rightarrow ... \rightarrow S_n$. Conditioned to a speech variety V_i and its set of rules R_i, it consists of finding those sequences of rule states q_i for the rule set r_i that leads to S_n. Therefore, the probability becomes:

$$P(S_n|B_m, V_i) = \sum_{q_i \in Q_i} P(B_m \xrightarrow{q_i} S_n) \qquad (10)$$

where Q_i is the set of valid sequences of rule states that transform the baseform B_m to the SV-specific form S_n for the given speech variety V_i, provided that at least one such sequence exists (otherwise the probability becomes zero). If the set Q_i is not empty, each term of the sum in equation 10 can be expressed as a product of probabilities of rules being in the required state to yield the output form S_n:

$$P(B_m \xrightarrow{q_i} S_n) = \prod_{j=1}^{L_i} P(q_i^j) \quad \text{iff } B_m \xrightarrow{q_i} S_n \qquad (11)$$

where L_i is the number of rules defined for the speech variety V_i. Furthermore, each rule state probability can be expressed as a function of the *a priori* rule probabilities (defined from knowledge-based sources):

$$P(q_i^j) = \begin{cases} 1 & \text{if } q_i^j = \text{state '0'} \\ P(r_i^j) & \text{if } q_i^j = \text{state '+'} \\ 1 - P(r_i^j) & \text{if } q_i^j = \text{state '-'} \end{cases} \qquad (12)$$

Finally, the probability associated to each selected SV-specific form S_n for a word W is the value of $P(S_n|W)$ with the appropriate substitutions seen above.

5. EXPERIMENTS

5.1. Database

All experiments were carried out on an internal telephone speech database called *Myosphere*. In this corpus, speakers from 12 speech varieties give a set of commands to a real speech recognizer (*e.g.*, "call Steve at office"). Most commands are short (3.8 words per sentence on average), but spontaneous and often uttered with hesitations and in different noisy conditions (background and line noise), so they represent fairly well a real life situation. Speech files include several annotations, including the speaker gender and his/her dominant speech variety.

5.2. Baseline system

A baseline HMM system was trained using HTK [6]. More than 90000 sentences uttered by more than 440 speakers were used for training. All 12 speech varieties were included, although around 80% of sentences were uttered by speakers of SAE or Northern Inland English, a dialect which largely shares SAE's phone inventory. Models consist of 39 monophones with 5 Gaussian mixtures per state, trained from 39 MFCC coefficients (12 static + 1 energy, 13 Δ, 13 $\Delta\Delta$). Additional models for silences and short pauses were also trained. As mentioned in section 3.3, no models specific to non-SAE SVs were used.

Five speech varieties were used for evaluation: Standard American English (SAE), Northern Inland English (NI), British English (Br), Indian English (In) and Asian-accented English (As). Eight to ten speakers (4-5 male, and 4-5 female) with an average of 164 sentences per speaker were used for each speech variety evaluation. A backoff bigram language model which was generated from all sentences of the database helped constrain the search[2]. Two different baseline lexica were used: the first lexicon (BLex1) contains only one baseform pronunciation per word, while the second lexicon (BLex2) is an expanded version of the first one with pronunciation variants created by phoneticians[3]. The vocabulary size for both lexica is 3815 words. Table 1 gives the baseline recognition results in WER. It shows that the WER may substantially increase with speech variety (*e.g.*, the Br WER is almost double the SAE WER).

5.3. Training of decision trees

As explained in section 3, a separate tree was trained for each speech variety and phone, with candidate transformations coming from all (12) SV-specific sets of rules applied to the baseform pronunciations (found in the BLex2 lexicon). Please note that due to a lack of data, trees for the As SV had to be trained from the test set as well, and therefore all tree-related results for the As SV reported in the next sections are for indication only. However, the SSA process itself strictly uses the adaptation set for all speech varieties.

SVs	SAE	NI	Br	In	As
Base (BLex1)	18.92	21.60	36.95	24.37	32.92
Base (BLex2)	18.31	20.92	34.93	23.45	31.31

Table 1. Baseline recognition results (% WER)

5.4. Results with SSA

The method of section 4 was applied to the whole adaptation set (140 sentences on average per speaker[4]). Before computing the SVP probabilities, SV-specific phone realizations that occurred less than 5 times were pruned to keep only reliable transformations. All speech varieties, words and baseforms used to compute the SVP probabilities were considered equiprobable ($P(V_i) = 1/N(V_i)$, $P(W_j) = 1/N(W_j)$ and $P(B_m) = 1/N(B_m)$) so that the final results are not biased towards any speech variety without any knowledge about the speaker's pronunciation characteristics. Some additional pruning thresholds – maximum 3 pronunciations per word, and stop when the sum of the highest output form probabilities equals or exceeds 0.7 – were also applied on the generated user lexicon to keep the lexicon small and to limit lexical confusability. Table 2 shows the results obtained. Unfortunately, our current implementation of the SVP concept showed very little improvement relative to the BLex2 baseline results.

SVs	SAE	NI	Br	In	As
Base (BLex2)	18.31	20.92	34.93	23.45	31.31
SVP (rules)	18.85	21.05	35.72	23.37	31.40
SVP (trees)	17.99	20.86	34.36	23.85	29.36

Table 2. Results with SSA (% WER)

5.5. Comparison with acoustic speaker adaptation (ASA)

The same experiments as mentioned in the previous subsection were carried out on ASA-adapted HMMs. The ASA method used was Maximum Likelihood Linear Regression (MLLR) with an 8-base regression class tree to cluster acoustically similar mixture components before evaluating the transformations (see [6] for more details). The amount of adaptation data was the same as for the SSA technique. Table 3 shows that application of ASA is much more effective than

[2]Test sentences were voluntarily included so that the OOV problem would not influence the results of our experiments.

[3]The average number of pronunciations per word in BLex2 is slightly higher than the CMU and BEEP dictionaries.

[4]This seems a large dataset, but since sentences are short they are equivalent to 30-35 sentences of Wall Street Journal (WSJ0) in terms of number of words.

SSA. However, we also notice that SSA performs better when combined with the ASA technique (up to +7.8% relative improvement with decision trees over the ASA baseline, +11.7% for the As SV). It seems that since lexica generated by SSA are speaker-dependent, they work better when the acoustic models are also speaker-dependent. SSA and ASA are applied at distinct levels of the system, and these results suggest that they are complimentary.

SVs	SAE	NI	Br	In	As
ASA+BLex2	11.74	12.96	20.59	13.91	19.04
ASA+SVP (rules)	12.13	12.82	20.53	14.18	19.18
ASA+SVP (trees)	11.04	12.48	18.99	13.26	16.81

Table 3. Results of SSA techniques over ASA (% WER)

5.6. Result analysis and suggestions for future work

In order to understand why the SSA experiments failed to bring more substantial improvement, results of Viterbi alignments were analyzed on the adaptation data. They show that:

1. Many pronunciations selected by the Viterbi alignments were associated with more than one speech variety: 78% with rules and 87% with trees (most of them were common to all 5 SVs).
2. Most of these selected and shared pronunciations were baseforms found in the BLex1 lexicon.

Given the first remark, the more speech varieties which accept a selected pronunciation as a possible SV-specific form, the more difficult it is to decide which speech variety best describes a speaker's pronunciation. The second remark tells us that since baseforms are most often preferred, the SVPs should be biased towards those speech varieties that most resemble SAE, namely SAE and NI. It is indeed the case with the knowledge-based method (rules) – SVP probabilities for SAE and NI are higher than for the other SVs, although not by much as noted in the first observation above (all SVP probabilities range between 0.15 and 0.24). Decision trees do not completely follow this assumption since preference for baseforms by a given SV is data-driven. They bias the SVPs slightly more towards the targeted SVs, but again the preference for one SV over another is not great.

Since baseforms are most often preferred, SV-specific variants added to the lexicon can increase lexical confusability more than they help with modeling pronunciation variation, which may explain the lack of improvement. According to our analysis, one or more of the following points may cause this preference for baseforms:

Database : Speakers were knowingly interacting with an ASR system and voluntarily spoke carefully so that their requests could be understood.

Phone inventory : The lack of phones specific to non-SAE SVs along with related acoustic models and rules (see section 3.3) biased the Viterbi alignment results towards the baseline SAE speech variety.

Acoustic models : Training of acoustic models was heavily biased towards the SAE and NI speech varieties due to the lack of data for the other SVs, which would bias the recognition results as well.

We believe that significant improvement can be achieved if the above issues are addressed.

6. CONCLUSIONS

In this paper we addressed the issue of modeling pronunciation variation of multiple speech varieties by introducing a method called *Symbolic Speaker Adaptation* (SSA). Although relatively little improvement has been obtained so far, the following points may still be useful for future experiments and should be retained:

- It is difficult to hone in on a speaker's speech variety when a high proportion of pronunciations selected by the Viterbi alignments during adaptation are common to all SVs.
- Though ASA performs better than SSA, their effect on system accuracy seems to be complimentary.
- Greater efficiency of SSA is expected if the database, phone inventory and acoustic model issues are addressed.

7. REFERENCES

[1] J.J. Humphries and P.C. Woodland, "Using accent-specific pronunciation modeling for improved large vocabulary continuous speech recognition", Eurospeech 97.

[2] K. Kumpf and R.W. King, "Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks", Eurospeech 97.

[3] J.P. Campbell, "Speaker recognition: A Tutorial", Proceedings of the IEEE, vol. 85, pp. 1437-1462, 1997.

[4] R. Carter and D. Nunan, "The Cambridge guide to teaching English to speakers of other languages", Cambridge University Press, eds. 2001, Cambridge.

[5] L. Breiman, J. Friedman, R. Olshen and C. Stone, "Classification and regression trees", Wadsworth International Group, 1984.

[6] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, "The HTK Book, Version 2.2", Cambridge University Engineering Department, 1999.