SYMBOLIC SPEAKER ADAPTATION WITH PHONE INVENTORY EXPANSION

Kyung-Tak Lee[1,2], *Lynette Melnar*[1], *Jim Talley*[1], *Christian J. Wellekens*[2]

Motorola Labs, Schaumburg (IL) & Austin (TX), USA[1]
Institut Eurécom, Sophia Antipolis, France[2]
{lee, wellekens}@eurecom.fr, {Lynette.Melnar, Jim.Talley}@motorola.com

ABSTRACT

This paper further develops a previously proposed adaptation method for speech recognition called Symbolic Speaker Adaptation (SSA). The basic idea of SSA is to model a speaker's pronunciation as a blend of speech varieties (SVs) - regional dialects and foreign accents - for which the system has existing pronunciation models. The system determines during an adaptation process the relative applicability of those models, yielding a speech variety profile (SVP) for each speaker. Speaker-dependent lexica for recognition are determined from a speaker's SVP. In this paper, we discuss a series of experiments designed to analyze how the SSA method is affected by SV-balanced training, expanded phone inventories, reduced amounts of adaptation data, and speech from SVs not modeled by the system. The most dramatic improvements were obtained by using expanded ("SV-inclusive") phone inventories. SSA was also shown to be effective with a very small number of adaptation sentences. And, SSA's SV blending scheme yields higher accuracy than using a SV classification scheme for speakers of novel (unseen) SVs.

1. INTRODUCTION

Several papers (*e.g.*, [1]) demonstrate that performance of an ASR system trained on a particular SV can significantly degrade when it is evaluated on another SV. It has been shown that pronunciation modeling methods help compensate in part for this increase in word error rate. An issue commonly addressed concerns modeling pronunciation variations when the non-standard SV is assumed to be known (*e.g.*, [2]). Although such methods effect good pronunciation modeling and contribute to performance improvement, they are limited to the targeted speech variety. A more difficult situation is when there are multiple SVs involved and the targeted SV is not known in advance. For such tasks, SV-specific pronunciation models may be combined with existing SV classification methods (*e.g.*, [3]) for multiple pronunciation targeting. However, these methods are designed to activate one single speech variety at a time. This one-SV classification scheme is inadequate in at least two respects: first, some speakers are best characterized by multiply modeled speech varieties (*e.g*, a bilingual or multilingual person), and second, in practice it is impossible to model all speech varieties of a given language and some speakers' SVs may fall outside the modeled pronunciation space covered by the system.

In [4], we introduced a method called Symbolic Speaker Adaptation (SSA) that addresses these shortcomings. It combines SV selection and pronunciation modeling and assumes - in contradistinction to general classification methods - that a speaker's pronunciation is best characterized not by a single SV modeled but rather a *combination* of them. This paper will describe some experiments and results that illustrate this assumption. Moreover, we will study the relative influence of the training database, phone inventory and the number of adaptation sentences available to our system.

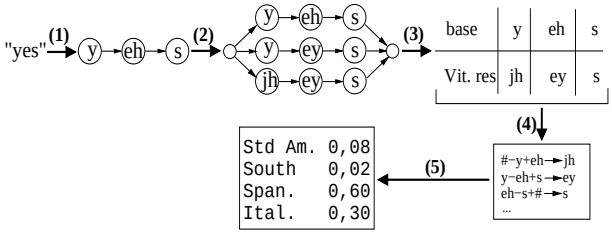

Fig. 1. Symbolic Speaker Adaptation

2. METHODOLOGY

Details about the SSA process can be found in [4]. This section briefly reviews some of its main aspects.

2.1. Overview

When a new speaker is enrolled in SSA, the system has no idea about his/her pronunciation characteristics, but it makes the assumption that he/she is well modeled by a combination of the speech varieties for which the system has existing pronunciation models. In order to model his/her pronunciation style, the speaker is enrolled in an adaptation process as depicted in Figure 1. The objective is to build a Speech Variety Profile (SVP) for this speaker. An SVP is simply a list of modeled SVs with their relative importance (probabilities) that best describe the speaker's pronunciation. 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 sets of pronunciation models (one set per SV), 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.
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.

The example in Figure 1 illustrates a possible adaptation scenario for a Spanish-accented English speaking person. Probabilities for the speaker profile depend on how frequently the speaker's pronunciations match the symbol transformations listed during the adaptation process and how accurately the same symbol transformations target the speech varieties modeled. The resulting SVPs influence then how a lexicon of baseforms (Standard American English (SAE) baseforms in our experiments) is filtered and transformed into a speaker specific set of pronunciation variants for use during recognition. The processes of adapting SVPs and generating user-specific lexica are explained in detail in [4].

2.2. Comparison to acoustic speaker adaptation

The basic concept of SSA is close to the CAT [5] and eigenvoice [6] techniques in Acoustic Speaker Adaptation (ASA): they form models of any speaker as a weighted sum of canonical speaker models (Gaussian means or eigenvoices). In a similar way, any speaker's speech variety(ies) can be represented as a point in the pronunciation space, and the objective of SSA is to find the coordinates of this point according to a set of basis vectors (represented by the different speech varieties modeled), or (ideally) the coordinates of its projection if the point is not located in the subspace spanned by the basis vectors. However, SSA is still different from the ASA techniques above because it does *not* alter the acoustic models, but only the lexicon, leaving the acoustic models truly speaker independent.

3. EXPERIMENTS

3.1. Database

All experiments were carried out on an internal English telephone speech database called *Myosphere*. In this corpus, speakers from 12 speech varieties give a set of commands to a real ASR system (*e.g.*, "call Steve at office"). Most commands are short (3.8 words per sentence on average), but spontaneous and in various noisy conditions (*e.g.*, cross-talk, line noise). Speech files include several annotations, including the speaker gender and his/her dominant speech variety.

3.2. Pronunciation models

For each speech variety (SV) and phone combination a decision tree was trained to predict SV-specific phone(s) from a canonical phone and its left and right contexts. For each training sentence, correspondences between the canonical phone transcription and its SV-specific phone transcription were derived using a Dynamic Programming (DP) based string alignment technique. For the canonical phone transcription, the word transcriptions are mapped, using the SAE lexicon, to an utterance pronunciation string. The SV-specific phone transcription is from the results of Viterbi selection of the best path through a network of recognition results. The pronunciation network for recognition is built from the baseform transcription(s) using some knowledge-based SV-specific sets of rules (see [4] for more details). The trees built from the canonical / SV-specific correspondences use questions related to phonetic features (*e.g.*, front, back, round, ...) for the immediate left and right contexts. The CART algorithm [7] was used to train the decision trees from the DP alignment results.

3.3. ASR systems

All ASR systems described in the following subsections are based on HMMs trained using HTK [8]. All models consist of phone-level monophones with 10 Gaussian mixtures per state, trained from 39 MFCC coefficients (12 static + 1 energy, 13 Δ, 13 $\Delta\Delta$). There are basically 41 distinct symbols, but this phone inventory was increased up to 164 to take account of the different speech varieties involved (more detail will follow). For evaluation, a back-off bigram language model was generated from all sentences of the database to help constrain the search[1]. The Standard American English (SAE) baseform lexicon contains 3815 words with pronunciation variants that are considered common to all speech varieties.

The following subsections will describe the various experiments carried out.

3.4. Influence of a SV-balanced training

In prior work ([4]), we speculated about the negative effects of the strongly unbalanced training data set (80% SAE and Northern Inland English (NI)). In order to see to what extent availability of non-SAE training data influences the recognition performance, two different sets of HMMs were trained. The first set (*SAE-only*) was trained using 14016 sentences of SAE data only, while the second set (*Multi-SV*) was trained using 14016 sentences evenly balanced (3504 sentences each) between Standard American English (SAE), Northern Inland English (NI) (*e.g.*, Chicago), British English (Br) and Indian English (In). Sentences used for evaluation were uttered by nine to ten speakers of each of these four SVs. Table 1 shows the baseline recognition results. It is not surprising that the Multi-SV HMMs outperform the SAE-only HMMs with speech varieties significantly distinct from SAE, namely Br and In. Also, as would be expected, using models trained with 75% non-SAE data (Multi-SV) rather than 100% SAE data (SAE-only) causes the WER for SAE test data to rise, but only moderately. Overall, the Multi-SV HMM system was clearly better. It was chosen as the baseline for the remainder of the paper (hereafter referred to as *Base 41* since it uses the original 41 phone set).

SVs	SAE	NI	Br	In
SAE-only	17.12	19.21	36.65	26.18
Multi-SV	17.97	19.35	25.68	21.94

Table 1. Baseline recognition results (WER) with single SV (SAE) training vs. Multi-SV training

3.5. Baseline SSA with SV-balanced models

The SSA process was applied with our baseline Base 41 (Multi-SV) HMMs using the whole adaptation set (153 sentences on average per speaker[2]) to see if the method would benefit from acoustic models trained on a SV-balanced training data. Consistent, but small, improvements relative to the non-SSA baseline results were obtained, as shown in Table 2. The improvements are generally better than those reported earlier ([4]) where models were trained with more, but considerably less SV-balanced, data.

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

[2] Equivalent to 30-35 sentences of Wall Street Journal (WSJ0) in terms of number of words.

	SAE	NI	Br	In
Base 41	17.97	19.35	25.68	21.94
SSA 41	17.77	18.92	24.73	21.89

Table 2. SSA results with SV-balanced training data (WER)

3.6. Influence of an SV-inclusive phone inventory

Next, we tested whether SSA is better able to hone in on a speaker's speech variety (or varieties) and contribute to performance improvement when the basic phone inventory with 41 symbols is augmented with more SV-specific phones. For this purpose, four additional sets of HMMs were trained with 70, 100, 130 and 164 symbols respectively. These are compared with the 41 symbol baseline set (Base 41) as described in section 3.4 above. The HMM model set with 164 symbols was obtained by training four subsets of 41 SV-specific models using each corresponding subset of 3504 SV-specific training sentences from section 3.4. The symbols (appropriately tagged for SV) were then simply combined at the end of training. The remaining sets (70, 100 and 130) were trained like the 164-set at the initial stage, but their HMM states were then clustered with 3 different threshold levels (yielding 70, 100 and 130 models) before the number of Gaussian mixtures in each state was increased. Separate pronunciation models were also built for each set of HMMs. To take account of the new phones introduced, each original phonetic transcription found in pronunciation networks created during the training of decision trees (cf. section 3.2) had four versions, each referring to one of the four subsets of SV-specific phones. Table 3 shows the recognition results for each "phone inventory - speech variety" pair *before* SSA was applied and Table 4 shows the results *after* the method was applied.

	SAE	NI	Br	In
Base 41	17.97	19.35	25.68	21.94
Base 70	17.41	19.23	26.47	21.95
Base 100	16.59	19.54	29.27	24.32
Base 130	16.57	18.78	33.52	26.26
Base 164	17.22	19.67	35.27	26.44
SV Dep.	17.22	18.59	23.39	24.40

Table 3. Baseline results with SV-inclusive (expanded) phone inventories (WER)

	SAE	NI	Br	In
SSA 41	17.77	18.92	24.73	21.89
SSA 70	17.60	18.78	24.73	21.91
SSA 100	16.27	19.26	24.53	22.84
SSA 130	17.70	18.64	23.89	25.83
SSA 164	17.70	19.00	23.13	26.73

Table 4. SSA results with SV-inclusive (expanded) phone inventories (WER)

Table 3 contains the ASR WER test results for each of the included SVs (columns) and for each of the trained HMM sets (rows). All tests were run using the standard (SAE) baseline lexicon which means that only SAE symbols were included in the recognition tests. To get an idea of the expected upper bound on performance, we also ran a "cheating experiment" for each SV, recognizing the test utterances for each SV with the 41 models trained exclusively on the 3504 training utterances for that SV. Those results are included as the last line in Table 3, labeled as "SV Dependent" results. Given the choice of the standard lexicon for the baseline experiments, the SAE SV dependent and the SAE Base 164 tests become one and the same (and they are similar to the Base SAE-only experiment of Table 1, but with one fourth the training data). We see that, for SAE, collapsing only the most similar phones (*e.g.*, in SAE Base 100 & 130) leads to improvement over the SAE SV dependent results, while forcing too much cross-SV collapsing causes WER increase relative to the SV dependent case. For the remaining three SVs, the addition of cross-SV data has a beneficial effect, with In even showing a 10% reduction in WER relative to its SV-dependent ("cheating") test.

Examination of the SSA results in Table 4 leads us to note that, with the exception of a few cases (mainly within SAE which was already well matched in the baseline), the SSA process consistently leads to WER reductions relative to the corresponding baseline (non-SSA) results. These can be dramatic (*e.g.*, by 34.4% error reduction in the case of Br 164). We also note that for each SV, there is at least one expanded inventory that produced as good or better WER as that yielded by the SSA process with the original minimal phone set (SSA 41). Compared to the Base 41 baseline, the improvement was best for the Br SV with 9.9% relative reduction in WER. Unfortunately, the relationship between phone set size and best WER for each SV is rather opaque.

The Indian English (In) results remain somewhat of an enigma. Large gains in performance were obtained by recognizing with multi-SV trained models, besting considerably even the SV-specific recognition results for In. Though the SSA does generally yield WER reductions for In relative to the corresponding non-SSA results, they are not nearly as dramatic as those for Br. And, rather than following the pattern of improving SSA results with larger sets of (more precise) models to recruit from, SSA performance decreases with model inventory size for In.

3.7. Influence of limited adaptation data

All of the SSA results presented thus far have been based upon utilizing the full adaptation data set for each test speaker (approx. 153 short utterances on average). Recall that in SSA the adaptation data is used to calculate an estimate of the SVP (Speech Variety Profile) for that speaker. The SVP characterizes the blend of the existing pronunciation models which will be used to form speaker specific pronunciation expectations (i.e., the speaker adapted lexicon). In these experiments, we examined the effect of reducing the available adaptation data on SVP estimation by using only five sentences for adaptation instead of 153. Table 5 gives the average probabilities of each modeled SV for the 10 Br test speakers (5 male, 5 female) using the 164 phone inventory. We see that, on average, using only 5 short adaptation utterances yields estimated SVPs which are virtually identical to those estimated with the full (153 utterance) adaptation sets. Tests across the variety of SVs and phone inventories led to similar results.

If we are modeling the SV phone inventories well, then we would expect a strong positive correlation between the accuracy of SVP identification and the accuracy of SSA-adapted ASR. The last column of Table 5 presents the average WERs obtained for the SVPs derived from the full set of (153) adaptation sentences and from only 5 sentences. We have found that the SSA method converges to a reasonable characterization of the SV of speakers

of modeled SVs with very little available data. Thus, it is suitable for tasks which require rapid adaptation. However, since SVP convergence is solely based on the set of actually occurring phone transformations, results will naturally be more reliable if larger quantities of adaptation data and / or phonetically balanced data are available.

	SAE	NI	Br	In	WER
153 sents	0.13	0.10	0.75	0.02	23.13
5 sents	0.11	0.11	0.74	0.04	23.20

Table 5. Average speaker SVP (Speech Variety Profile) probabilities and WERs for the British English test speakers

3.8. Comparison between SV classification and SV blending schemes

We also investigated system performance with non-modeled speech varieties by comparing two different schemes: a *classification* scheme that only selects the best SV in adaptation derived SVPs, and the SSA *SV blending* scheme that keeps all SVs with their respective probabilities. To have a fair comparison, the number of pronunciation variants used for each speaker were made to be approximately the same for both schemes. Fourteen speakers of a diverse group of non-modeled SVs were evaluated. There were two regional / dialectal varieties: African-American (Af), one speaker, and two American speakers with Southern accents (So). Additionally, three varieties of foreign accented English were represented: Spanish (Sp), two speakers, Asian (As), seven speakers, and German (Ge), two speakers. Results using the full phone inventory set are given in Table 6 and are structured as follows:

Class. shows the results obtained with a classification scheme along with the selected SV in parentheses.

Ideal shows the results obtained with an *ideal* classifier (or an oracle) that always selects the SV that leads to the lowest WER, along with the corresponding SV in parentheses.

SSA shows the results obtained with the SV blending scheme.

In the last column, all WERs equal to or lower than the matching classifier scheme counterparts are marked in bold, and those among them that are equal to or lower than the "ideal" classifier WERs are further marked with a '*'. We observe that SSA performs on average better than a classification scheme method (7.2% relative improvement) and is comparable to the "ideal" classifier. Similar behavior can be observed with lower phone inventory HMMs.

	Class.		Ideal		SSA
Af	13.17	(NI)	12.18	(SAE)	**12.82**
As (1)	9.78	(NI)	9.06	(SAE)	10.51
As (2)	26.76	(NI)	26.76	(NI)	27.73
As (3)	47.69	(Br)	30.00	(NI)	**31.89**
As (4)	51.42	(NI)	51.42	(NI)	**47.17 ***
As (5)	34.67	(SAE)	34.67	(SAE)	**33.33 ***
As (6)	32.14	(Br)	32.14	(Br)	**30.71 ***
As (7)	34.63	(SAE)	34.63	(SAE)	**33.98 ***
Ge (1)	51.52	(Br)	51.52	(Br)	**51.52 ***
Ge (2)	40.96	(Br)	38.55	(SAE)	**40.96**
So (1)	35.98	(NI)	35.98	(NI)	**35.15 ***
So (2)	5.45	(SAE)	3.96	(NI)	**5.45**
Sp (1)	34.15	(SAE)	34.15	(SAE)	**29.27 ***
Sp (2)	38.33	(Br)	26.67	(SAE)	**33.33**
Average	32.62		30.12		**30.27**

Table 6. Comparative results (WER) for handling unmodeled SVs between SVP based classification, the ideal classifier, and SSA's SV blending methods

4. CONCLUSION

In this paper, we explored different factors influencing the capability of SSA and showed that the addition of an SV-inclusive phone inventory may substantially contribute to improved performance. We also observed that only a few adaptation sentences are required for SSA and that a SSA's SV blending scheme is more appropriate than a classification scheme to generalize the use of the system to speakers of any SV of the same language.

A possible extension of this research line would be to explore the selection of an optimal SV set to model. For example, it is obvious from this paper that modeling SAE and NI separately is rather redundant because of their similar acoustic and phonetic properties. We believe that a data-driven approach to determine an optimal set of "canonical bases" that are not necessarily bound to any particular SV (using, for example, a concept similar to [6] at the SV level) could be advantageous.

5. REFERENCES

[1] C. Huang, E. Chang, J. Zhou and K.-F. Lee, "Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition, ICSLP 2000.

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

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

[4] K.-T. Lee, L. Melnar and J. Talley, "Symbolic speaker adaptation for pronunciation modeling", ITRW-PMLA 2002.

[5] M.J.F. Gales, "Cluster adaptive training for speech recognition", ICSLP 98.

[6] R. Kuhn, P. Nguyen, J.C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field and M. Contolini, "Eigenvoices for speaker adaptation", ICSLP 98.

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

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