INTRODUCING ATTRIBUTE FEATURES TO FOREIGN ACCENT RECOGNITION

Hamid Behravan[1], *Ville Hautamäki*[1], *Sabato Marco Siniscalchi*[2,3], *Tomi Kinnunen*[1] *and Chin-Hui Lee*[3]

[1]School of Computing, University of Eastern Finland, Finland
[2]Faculty of Architecture and Engineering, University of Enna "Kore", Italy
[3]School of ECE, Georgia Institute of Technology, USA

ABSTRACT

We propose a hybrid approach to foreign accent recognition combining both phonotactic and spectral based systems by treating the problem as a spoken language recognition task. We extract speech attribute features that represent speech and acoustic cues reflecting foreign accents of a speaker to obtain feature streams that are modeled with the i-vector methodology. Testing on the Finnish Language Proficiency exam corpus, we find our proposed technique to achieve a significant performance improvement over the state-of-the-art systems using only spectral based features.

Index Terms— Speech attributes, i-vector, foreign accent recognition, language recognition

1. INTRODUCTION

In *automatic foreign accent recognition*, we aim to detect speaker's mother tongue (L1) when he or she is speaking in another language (L2) [1]. When speaking in L2, the speaker's accent is usually colored by the learned patterns in L1 [2]. When the native language is spoken instead, it can be said to vary in terms of its regional dialects and accents. *Dialect* refers to linguistic variations of a language, while *accent* refers to different ways of pronouncing a language within a community [3]. In the NIST *language recognition evaluation* (LRE) scenarios, dialect and accent recognition have been included as sub-tasks. As an example, the most recent LRE 2011 covered four different Arabic dialects as target languages [4]. Foreign accent recognition, however, differs from common accent recognition in two major distinctions. Firstly, non-native speaker's *accentedness* partly depends on the language proficiency [2]. Secondly, the L2 is a noisy channel through which the identity of the mother tongue is transmitted.

In this study we treat foreign accent recognition as a language recognition task typically accomplished via either *acoustic* or *phonotactic* modeling [5]. In the former approach, acoustic features, such as *shifted delta cepstra* (SDC), are used with bag-of-frames models, such as *universal background model* (UBM) with adaptation [6, 7]. The latter is based on the hypothesis that dialects or accents differ in terms of their phone sequence distributions. It uses phone recognizer outputs, such as n-gram statistics, together with a language modeling back-end [8, 9].

Among the choices for acoustic modeling, the recent *i-vector* paradigm [10] has proven successful in both speaker [10, 11], language [12], and accent recogntion [13]. It extracts a low-dimensional representation of the sequence of feature vectors. Session and channel variability is typically tackled with techniques such as *linear discriminant analysis* (LDA). The i-vectors from spectral features have been used in dialect and foreign accent characterization. In [14], L1 of the non-native English speakers was recognized using multiple spectral systems, including i-vectors with different back-ends. The i-vector based system outperformed other compared methods most of the time. In [1], it was found out that the i-vector system using SDCs outperformed other methods in recognizing Finnish non-native accents.

In language recognition, spectral features with i-vector based systems have been seen to outperform the classical phonotactic language recognition [4]. However, *knowledge based* modeling, such as phonotactic features, are known to be linguistically and phonetically relevant [5]. However, the front-end of the phonotactic system needs a tokenizer that will turn the utterance into a sequence of "phonetic letters" [15, 16]. An ad-hoc approach is to use a phone recognizer developed for one language, such as Hungarian, and apply it to all phonotactic recognition tasks [17].

In the present work, we argue that, especially in foreign accent recognition, a universal phonetic tokenizer is preferable. It will be able to find differences between the unknown L1 and the known L2. For example, Spanish L1 speaker trying to pronounce Finnish word "*stressi*" (stress) will typically lead to /e/ placed as a prefix, leading to "*estressi*". In this case, detecting a vowel in the beginning of the word is a cue for Spanish L1. We then propose to use speech attributes [18, 19, 20] to represent a language-universal set of units to be modeled. In addition, we avoid the early quantization of the attribute detector scores by computing an i-vector from the detector score vector streams.

This work was partially supported by Academy of Finland (projects 253000 and 253120).

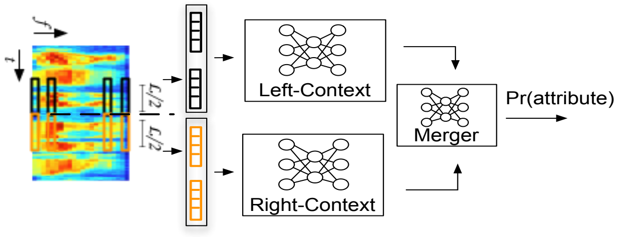

Fig. 1. The internal structure of an attribute detector is shown. Energy trajectories are fed into the left-context and right-context ANNs. A merger then combines the outputs generated by those two neural networks and produced the final attribute posterior probabilities.

2. SPEECH ATTRIBUTE EXTRACTION

2.1. Choice and Extraction of Attribute Features

The set of speech attributes used in this work is mainly acoustic phonetic features, and it comprises five manner of articulation classes (**glide, fricative, nasal, stop, and vowel**), and **voicing**. Those attributes could be identified from a particular language and shared across many different languages, so they could also be used to derive a universal set of speech units. Furthermore, data-sharing across languages at the acoustic phonetic attribute level is naturally facilitated by using these attributes, so more reliable language-independent acoustic parameter estimation can be anticipated [21]. In [16], it was also shown that these attributes can be used to compactly characterize any spoken language along the same lines as in the automatic speech attribute transcription (ASAT) paradigm for automatic speech recognition (ASR) [20]. Therefore, we believe that it can also be useful to characterize speaker accent.

Data-driven detectors are used to spot speech cues embedded in the speech signal. An attribute detector converts an input utterance into a time series that describes the level of presence (or level of activity) of a particular property of an attribute over time. A bank of six detectors is used in this work, each detector is individually designed for spotting of a particular event. Each detector is realized with three single hidden layer feed-forward ANNs (artificial neural networks) organized in a hierarchical structure and trained on sub-band energy trajectories that are extracted with a 15 band uniform mel-frequency filterbank. For each critical band a window of 310ms centered around the frame being processed is considered and split in two halves: left-context and right-context [22]. Two independent front-end ANNs ("lower nets") are trained on those two halves and generate left- and right-context speech attribute posterior probabilities, respectively. The outputs of the two lower nets are then sent to the third ANN that acts as a merger and gives the attribute-state posterior probability of the target speech attribute. Figure 1 shows the detector architecture in detail.

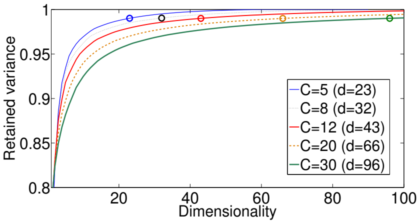

Fig. 2. Remaining variance after PCA. Comparing stacked context sizes 5, 8, 12, 20 and 30 frames.

2.2. Long-term Attribute Extraction

Each attribute detector outputs probabilities $p(H^{(i)}_{\text{target}}|\boldsymbol{f})$, $p(H^{(i)}_{\text{anti}}|\boldsymbol{f})$ and $p(H^{(i)}_{\text{noise}}|\boldsymbol{f})$, of target class i, non-target and noise model, given a speech frame $\boldsymbol{f}$. All these probabilities sum to one. We then form a new feature vector $\boldsymbol{x}$ by concatenating each of these posteriors of each six target classes. Since language and dialect recognizers benefit from an inclusion of long temporal context, it is natural to study similar ideas for attribute modeling. The first idea is to compute SDCs from the attribute features, treating them analogous to cepstral coefficients. But since this is difficult to interpret, we study a simple feature stacking. To this end, let $\boldsymbol{x}(t)$ denote the 18-dimensional (6 attributes × 3 features) attribute vector at frame t. We form a sequence of new $p = 18 \times C$ dimensional stacked vectors $\tilde{\boldsymbol{x}}_C(t) = (\boldsymbol{x}(t)^*, \boldsymbol{x}(t+1)^*, \ldots, \boldsymbol{x}(t+C-1)^*)^*$, $t = 1, 2, \ldots$, where C is the context size and * stands for transpose. Principal component analysis (PCA) is used to project each $\tilde{\boldsymbol{x}}_C(t)$ onto the first $d \ll p$ eigenvectors corresponding to the largest eigenvalues of the sample covariance matrix. We estimate the PCA basis from the same data as the UBM and the T-matrix, after VAD.We set d to retain 99 % of the cumulative variance. As Fig. 2 indicates, d varies from ∼20 to ∼100, with larger dimensionality assigned to longer context as one expects.

3. RECOGNIZING FOREIGN ACCENTS

3.1. I-vector Modeling

We now shortly review i-vector extraction. It is grounded on the *universal background model* (UBM), which is a M-component Gaussian mixture model parametrized by $\{w_m, \boldsymbol{m}_m, \boldsymbol{\Sigma}_m\}, m = 1, \ldots, M$, where we have mixture weight, mean vector and covariance matrix, respectively. We here restrict the covariance matrix to be diagonal. The i-vector model is defined for the UBM component m as [10]:

$$\boldsymbol{s}_m = \boldsymbol{m}_m + \boldsymbol{V}_m \boldsymbol{y} + \epsilon_m, \tag{1}$$

where $\boldsymbol{V}_m$ is the sub-matrix of the total variability matrix, $\boldsymbol{y}$ is the latent vector, called an i-vector, ϵ_m is the residual term and $\boldsymbol{s}_m$ is the m'th sub-vector of the utterance dependent supervector. The ϵ_m is distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_m)$, where

$\boldsymbol{\Sigma}_m$ is a diagonal matrix. Given all these definitions, posterior density of the $\boldsymbol{y}$, given the sequence of observed feature vectors, is Gaussian. Expectation of the posterior is the extracted i-vector. Hyperparameters of the i-vector model, $\boldsymbol{m}_m$ and $\boldsymbol{\Sigma}_m$ are copied directly from UBM and $\boldsymbol{V}_m$ are estimated by EM algorithm from the same corpus as is used to estimate the UBM.

3.2. Scoring against Accent Models

We use *cosine scoring* [23] between two i-vectors $\boldsymbol{y}_{\text{test}}$ and $\boldsymbol{y}_{\text{target}}$ to match test utterance to target L2 language model. Cosine score is given by the dot product $\langle \hat{\boldsymbol{y}}_{\text{test}}, \hat{\boldsymbol{y}}_{\text{target}} \rangle$,

$$\text{score}(\boldsymbol{y}_{\text{test}}, \boldsymbol{y}_{\text{target}}) = \frac{\hat{\boldsymbol{y}}_{\text{test}}^{\text{T}} \cdot \hat{\boldsymbol{y}}_{\text{target}}}{\|\hat{\boldsymbol{y}}_{\text{test}}\| \, \|\hat{\boldsymbol{y}}_{\text{target}}\|}, \tag{2}$$

where $\boldsymbol{A}$ is the HLDA projection matrix trained by using all training utterances and $\hat{\boldsymbol{y}}_{\text{test}}$ is,

$$\hat{\boldsymbol{y}}_{\text{test}} = \boldsymbol{A}^{\text{T}} \boldsymbol{y}_{\text{test}}. \tag{3}$$

In order to model $\hat{\boldsymbol{y}}_{\text{target}}$, we followed the same strategy used in [4], where $\hat{\boldsymbol{y}}_{\text{target}}$ is defined as

$$\hat{\boldsymbol{y}}_{\text{target}} = \frac{1}{N_d} \sum_{i=1}^{N_d} \hat{\boldsymbol{y}}_{id}, \tag{4}$$

where N_d is the number of training utterances in dialect d, and $\hat{\boldsymbol{w}}_i$ is the projected i-vector of training utterance i for accent d computed the same way as in (3).

4. EXPERIMENTAL SETUP

4.1. Corpora

The "stories" part of the OGI Multi-language telephone speech corpus [24] was used to train the articulatory detectors. This corpus has phonetic transcriptions for six languages: English, German, Hindi, Japanese, Mandarin, and Spanish. Data from each language were pooled together to obtain: 5.57 hours for the training set, and 0.52 hours for the validation set.

A series foreign accent recognition experiments was performed on the *FSD* corpus [25] which was developed to assess Finnish language proficiency among adults of different nationalities. These selected the oral responses portion of the exam, corresponding to 18 foreign accents. Since the number of utterances is small, 9 accents — Russian, Albanian, Arabic, Chinese, English, Estonian, Kurdish, Spanish, and Turkish — with enough available data were used. The unused accents are, however, used in training the UBM and the V_m-matrices. For our purposes, each accent set is randomly split into a test and a train set. The test set consists of (approximately) 30% of the utterances, while the training set consists of the remaining 70% to train foreign accent recognizers. The raw audio files were partitioned into 30 sec chunks and re-sampled to 8 KHz. Statistics of the test and train portions are shown in Table 1.

Table 1. Train and test file distributions in the FSD corpus.

Accent	#train files	#test files	#speakers
Spanish	60	25	15
Albanian	67	30	19
Kurdish	83	35	21
Turkish	84	34	22
English	92	37	23
Estonian	153	63	38
Arabic	166	67	42
Russian	599	211	235

Table 2. Sliding window context experiments with PCA as a dimensionality reduction.

PCA features	Pooled EER (%)	$C_{\text{avg}} \times 100$
$(C = 5, d = 23)$	10.65	4.82
$(C = 20, d = 50)$	10.44	4.71
$(C = 30, d = 96)$	**8.73**	**4.47**

4.2. Attribute Detector Design

One-hidden-layer feed forward multi-layer perceptrons (MLPs) were used to implement each attribute detector shown in Figure 1. The number of hidden nodes with a sigmoidal activation function is 500. MLPs were trained to estimate attribute posteriors, and the training data were separated into "feature present," "feature absent," and "other" regions for every phonetic class used in this work. The classical back-propagation algorithm with a cross-entropy cost function was adopted to estimates the MLP parameters. To avoid over-fitting, the reduction in classification error on the development set was adopted as the stopping criterion. The attribute detectors employed in this work were actually just those used in [21].

4.3. Evaluation Protocol

System performance is reported in terms of *equal error rate* (EER) and average detection cost (C_{avg}) [5]. Results are reported per each accent for a cosine scoring classifier. C_{avg} is defined as [5],

$$C_{\text{avg}} = \frac{1}{J} \sum_{j=1}^{M} C_{\text{DET}}(L_j), \tag{5}$$

where $C_{\text{DET}}(L_j)$ is the detection cost for subset of test segments trials for which the target accent is L_j and J is the

Table 3. Summary of results and compared against baseline spectral system, results are shown in pooled EER and C_{avg}.

Features (dimensionality)	Pooled EER (%)	$C_{\mathrm{avg}} \times 100$
SDC+MFCC(56)	15.00	7.00
Attribute(18)	12.54	5.07
Attribute+Δ(36)	11.33	4.79
Attribute+Δ+$\Delta\Delta$(54)	11.00	4.59
PCA features(96)	**8.73**	**4.47**

number of target languages. The per target accent cost is then,

$$\begin{aligned} C_{\mathrm{DET}}(L_j) &= C_{\mathrm{miss}} P_{\mathrm{tar}} P_{\mathrm{miss}}(L_j) \\ &+ C_{\mathrm{fa}}(1-P_{\mathrm{tar}}) \frac{1}{J-1} \sum_{k \neq j} P_{\mathrm{fa}}(L_j, L_k). (6) \end{aligned}$$

The miss probability (or false rejection rate) is denoted by P_{miss}, i.e., a test segment of accent L_i is rejected as being in that accent. On the other hand $P_{\mathrm{fa}}(L_i, L_k)$ denotes the probability when a test segment of accent L_k is accepted as being in accent L_i. It is computed for each target/non-target accent pairs. Measures, C_{miss} and C_{fa}, are costs of making errors and both were set to 1. P_{tar} is the prior probability of a target accent and was set to 0.5.

5. EXPERIMENTS AND RESULTS

First we experimented with different context sizes ($C = 5, 20, 30$). Feature vectors were concatenated and PCA dimensionality reduction was trained on the held out data. Output dimensionality (d) was set to retain 99% percent of the cumulative variance. In Table 2 we see that increasing the context size from 5 to 30 will decrease the both pooled EER and C_{avg}. We also attempted to use context as large as 40 frames, which resulted to a numerical problems in UBM computation. Output dimensionality of 124 was too large with respect to the available data, so we observed singular Gaussian components.

We applied the context size 30 to the following experiments (see Table 3). We contrasted the above mentioned system to the baseline SDC+MFCC based system in [1]. In addition to sliding window based context modeling, we also employ standard Δ and $\Delta\Delta$ to attribute feature vectors. We notice that increasing the context size using Δ and $\Delta\Delta$ features improves marginally over not using the context at all. A large 30-frame context brought forth an improvement. All systems based on speech attributes improved substantially over the baseline. In Table 4 we show the per target accent error rates, in EER and C_{DET}. We notice that there is a large variation in error rates, where Turkish and Albanian are easiest and Russian and Estonian are the hardest to recognize.

Table 4. Per-language results for PCA features (30,96). The results are given in EER and C_{DET}.

Features	EER (%)	$C_{\mathrm{DET}} \times 100$
Spanish	9.00	4.10
Turkish	3.82	2.01
Albanian	4.34	2.48
English	8.11	4.20
Arabic	7.46	4.04
Russian	15.54	8.17
Kurdish	8.57	4.67
Estonian	12.70	6.11

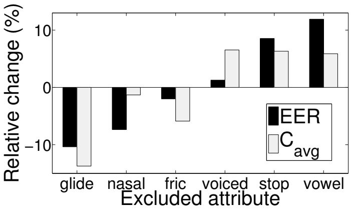

Fig. 3. Exclusion experiment, where relative change is shown when one attribute is left out.

We also studied the relative importance of individual speech attributes to system performance in Fig. 3. No context was used, so raw pooled EER is 12.54%. We left out one by one all attributes, so we had 15-dimensional feature vectors. We noticed that voicing, stop and vowels are individually beneficial (leaving any one of them out will decrease the system performance). On the other hand, glide, nasal and fricative are not individually useful. We also noticed that in terms of conclusions, pooled EER and C_{avg} agree. Usefulness of vowels in contrast to other features can be explained by the fact that Finnish has a very large vowel space (with 8 vowels) including vowel lengthening. It can create difficulties for L2 speakers to hit the correct vowel target, thus showing the L1 influence.

6. CONCLUSION

We proposed speech attributes as features for foreign accent recognition. Instead of using speech attributes directly in a phonotactic system, we modeled the sequence of speech attribute feature vectors using the i-vector methodology. The key idea is to treat foreign accent recognition as a language recognition task and use universal speech attributes. Speech attributes are employed because their statistics can differ considerably from one language to another. Indeed, all attribute feature configurations improved over the spectral-only baseline system. Moreover, adding context information allowed substantially better results. So far, we have only used manner of articulation features, yet place of articulation can further enhance accent recognition performance, as shown in [16]. As a future work, experiments on English foreign accent recognition will be carried out. Furthermore, the possible beneficial effect of combining SDC- and attribute-based information will be investigated.

7. REFERENCES

[1] H. Behravan, V. Hautamäki, and T. Kinnunen, “Foreign accent detection from spoken finnish using i-vectors,” in *Interspeech*, Lyon, France, August 2013.

[2] J. Flege, C. Schirru, and I. MacKay, “Interaction between the native and second language phonetic subsystems,” *Speech Communication*, vol. 40, no. 4, pp. 467–491, 2003.

[3] J. Nerbonne, “Linguistic variation in computation,” in *EACL*, Budabest, Hungary, 2003, pp. 3–10.

[4] E. Singer, P. Torres-Carrasquillo, D. Reynolds, A. McCree, F. Richardson, N. Dehak, and D. Sturim, “The MITLL NIST LRE 2011 language recognition system,” in *Speaker Odyssey*, Singapore, 2012, pp. 209–215.

[5] H. Li, K. A. Lee, and B. Ma, “Spoken language recognition: From fundamentals to practice,” *Proceedigns of the IEEE*, vol. 101, no. 5, pp. 1136–1159, May 2013.

[6] P. Torres-Carrasquillo, T. Gleason, and D. Reynolds, “Dialect identification using Gaussian mixture models,” in *Speaker Odyssey*, Toledo, Spain, 2004, pp. 757–760.

[7] G. Liu and J. H. Hansen, “A systematic strategy for robust automatic dialect identification,” in *EUSIPCO*, Barcelona, Spain, 2011, pp. 2138–2141.

[8] M.A. Zissman, T.P. Gleason, D.M. Rekart, and B.L. Losiewicz, “Automatic dialect identification of extemporaneous conversational latin american spanish speech,” in *ICASSP*, Detroit, USA, 1995.

[9] T. Wu, J. Duchateau, J. Martens, and D. Compernolle, “Feature subset selection for improved native accent identification,” *Speech Communication*, pp. 83–98, 2010.

[10] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech and Language Processing*, pp. 788–798, 2011.

[11] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, “I-vector based speaker recognition on short utterances,” in *Interspeech*, Florence, Italy, 2011, pp. 2341–2344.

[12] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, “Language recognition in i-vectors space,” in *Interspeech*, Florence, Italy, 2011, pp. 861–864.

[13] A. DeMarco and S. J. Cox, “Iterative classification of regional British accents in i-vector space,” in *Proc. Machine Learningin Speech and Language Processing*, Portland, USA, 2012.

[14] M.H. Bahari, R. Saeidi, H. Van hamme, and D. van Leeuwen, “Accent recognition using i-vector, Gaussian mean supervector, Gaussian posterior probability for spontaneous telephone speech,” in *ICASSP*, Vancouver, Canada, 2013.

[15] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, “Exploring universal attribute characterization of spoken languages for spoken language recognition,” in *Interspeech*, Brighton, UK, 2009, pp. 168–171.

[16] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, “Universal attribute characterization of spoken languages for automatic spoken language recognition,” *Computer Speech & Language*, vol. 27, no. 1, pp. 209–227, 2013.

[17] K. A. Lee, C. H. You, V. Hautamäki, A. Larcher, and H. Li, “Spoken language recognition in the latent topic simplex,” in *Interspeech*, Florence, Italy, 2011, pp. 2893–2896.

[18] C.-H. Lee, “From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition,” in *Interspeech*, Jeju Island, Korea, 2004, pp. 109–112.

[19] I. Bromberg, Q. Fu, J. Hou, J. Li, C. Ma, B. Matthews, A. Moreno-Daniel, J. Morris, S. M. Siniscalchi, Y. Tsao, and Y. Wang, “Detection-based ASR in the automatic speech attribute transcription project,” in *Interspeech*, Antwerp, Belgium, 2007, pp. 1829–1832.

[20] C.-H. Lee and S. M. Siniscalchi, “An information-extraction approach to speech processing: Analysis, detection, verification, and recognition,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089–1115, 2013.

[21] S. M. Siniscalchi, D.-C. Lyu, T. Svendsen, and C.-H. Lee, “Experiments on cross-language attribute detection and phone recognition with minimal target specific training data,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 875–887, 2012.

[22] P. Schwarz, P. Matějaka, and J. Černocký, “Hierarchical structures of neural networks for phoneme recognition,” in *ICASSP*, Toulouse, France, 2006, pp. 325–328.

[23] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, “Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verication,” in *Interspeech*, Brighton, UK, 2009, pp. 1559–1562.

[24] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, “The OGI multi-language telephone speech corpus,” in *ICSLP*, Banff, Canada, Oct. 1992, pp. 895–898.

[25] “Finnish national foreign language certificate corpus,” http://yki-korpus.jyu.fi.