Dialect Levelling in Finnish: A Universal Speech Attribute Approach

Hamid Behravan[1,4]*, Ville Hautamäki*[1]*, Sabato Marco Siniscalchi*[2,5]*, Elie Khoury*[3]*, Tommi Kurki*[4]*, Tomi Kinnunen*[1] *and Chin-Hui Lee*[5]

[1]School of Computing, University of Eastern Finland, Finland
[2]Faculty of Architecture and Engineering, University of Enna "Kore", Italy
[3]Idiap Research Institute, Switzerland
[4]School of Languages and Translation Studies, University of Turku, Finland
[5]School of ECE, Georgia Institute of Technology, USA

behravan@cs.uef.fi

Abstract

We adopt automatic language recognition methods to study dialect levelling — a phenomenon that leads to reduced structural differences among dialects in a given spoken language. In terms of dialect characterisation, levelling is a nuisance variable that adversely affects recognition accuracy: the more similar two dialects are, the harder it is to set them apart. We address levelling in Finnish regional dialects using a new SAPU (Satakunta in Speech) corpus containing material from Satakunta (South-Western Finland) between 2007 and 2013. To define a compact and universal set of sound units to characterize dialects, we adopt speech attributes features, namely manner and place of articulation. It will be shown that speech attribute distributions can indeed characterise differences among dialects. Experiments with an i-vector system suggest that (1) the attribute features achieve higher dialect recognition accuracy and (2) they are less sensitive against age-related levelling in comparison to traditional spectral approach.

Index Terms: Finnish language, dialect levelling, social factors, native dialects, dialect detection, attributes detection, i-vector modelling.

1. Introduction

Dialect refers to linguistic variations of a standard spoken language [1]. Over the years, stereotypical differences among dialects of the same spoken language have become smoother and smoother [2] due several co-occurring factors such as language standardisation, industrialisation (increased people mobility) and modernisation (mass media diffusion) [2, 3]. The reduction of peculiar differences among dialects is referred to as *levelling* [4, 5]. Levelling is a common phenomenon in languages. For example, the effect of levelling due to language standardisation can be seen in the phoneme /d/ that is a standard variant in Finnish and also in Pori and Rauma dialects. It has dialectal phonemes in all other dialects of this region, but shows levelling for instance in Honkilahti. That is, Honkilahti has been influenced by the standard Finnish.

In fact, spoken sentences produced by speakers of regional dialects may still be characterised by dialect-specific cues, but levelling weakens such cues, making automatic dialect recognition a hard task. In our task of dialect characterisation, we consider levelling to be a nuisance factor to be compensated for. The problem is analogous to foreign accent recognition [6, 7, 8, 9], where the speakers's second language (L2) masks his mother tongue (L1).

Automatic dialect recognition is traditionally treated as a language recognition problem. State-of-the-art language recognition techniques, either *acoustic* [10, 11] or *phonotactic* [12, 13, 14] ones, can be applied to regional dialect recognition [15, 16]. Although the former techniques have recently proven to attain better language recognition performance [17] by embedding acoustic spectral features within the i-vector framework, there are linguistic and paralinguistic cues (e.g., speaker's age, vocal tract articulators) which can be used for dialect discrimination. We, therefore, propose an articulatory-motivated features with an i-vector method. More specifically, so-called *automatic speech attribute transcription* (ASAT) approach [18, 19, 20, 21, 22] is adopted in order to generate the features of interest for this work, and a bank of detectors is built to detect the presence of speech attributes in a given segment of speech. The speech attributes chosen represent a language-universal set of units, namely manner and place of articulation classes, detected with the help of artificial neural networks.

Indeed, we have already demonstrated that by coupling universal attribute detectors and a state-of-the-art i-vector approach, Finnish foreign accents can be accurately discriminated [7]. Furthermore, ASAT speech attributes have been proven useful in automatic language recognition tasks [23] and cross-language recognition of "unseen" languages using minimal training data from the target languages [24]. The universality of our speech attributes can be better appreciated by thinking of that our detectors were *not* built using ad-hoc Finnish material. In fact, the set of attribute detectors is one used to carry out the independent language recognition experiments reported in [24].

A recently-collected SAPU (Satakunta in Speech) corpus is used to validate our approach. The SAPU corpus includes 8 Finnish sub-dialects or regional dialects and hundreds of speakers. The SAPU Corpus was collected in an interview setting, where subjects interacted with the interviewer in a conversational way. However, interviewer's speech is included in the recording, so needed to be removed by using speaker diarization.

We study three levelling factors: age, gender and place of birth. We first investigate how levelling affects dialect recognition accuracy. Then, the strength of levelling as a function of the speaker's age is investigated. We hypothesize that younger speakers might have lost some of the stereotypical features of their regional dialect, which might still be clear in older speakers of the same region.

2. System description

2.1. Attribute detection

The collection of information embedded into the speech signal, referred to as attributes of speech, also includes the speaker profile encompassing gender, accent, emotional state and other speaker characteristics, which may come useful to automatically uncover the speaker's dialect in a spoken utterance. Indeed, speakers from different regions of a same country may pronounce/produce nasal sounds with diverse acoustic characteristics. Moreover, speakers may also use speech patterns (i.e., conventions of vocabulary, pronunciation, grammar, and usage of certain words) that differ from region to region of the same nation. In this work, the speech attributes of interest are mainly phonetic features, and a bank of speech attribute detectors is built to automatically extract phonetic information from the speech signal. Specifically, five manner of articulation classes (**glide, fricative, nasal, stop, and vowel**), nine place of articulation classes (**coronal, dental, glottal, high, labial, low, mid, retroflex, and velar**), and **voicing** are used. 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, and reliable language-independent acoustic parameter estimation can be anticipated [24].

Each detector is individually designed for modelling a particular speech attribute, and it is built employing three single hidden layer feed-forward multi-layer perceptrons (MLPs) hierarchically organised as described in [25]. These detectors are trained on sub-band energy trajectories that are extracted with a 15 band uniform Mel-frequency filterbank. For each critical band a window of $310ms$ centred around the frame being processed is considered and split in two halves: left-context and right-context [26]. Two independent front-end MLPs ("lower nets") are designed on those two halves and deliver left- and right-context speech attribute posterior probabilities, respectively. Usually, the discrete cosine transform is applied to the input of these lower nets to reduce dimensionality. The outputs of the two lower nets are then sent to the third MLP that acts as a merger and gives the attribute-state posterior probability of the target speech attribute.

Overall, each detector converts an input speech signal into a time series which describes the level of presence (or level of activity) of a particular property of an attribute, or event, in the input speech utterance over time. By using MLPs, the posteriori probability of the particular attribute, given the speech signal, is computed. Articulatory detectors are trained using the same corpus as in [7].

2.2. I-vector modelling

I-vector modelling is rooted on Bayesian *factor analysis* technique which forms a low-dimensional *total variability space* containing both speaker and channel variabilities [27]. In this approach, *universal background model* (UBM), which is a M-component Gaussian mixture model parameterised by $\{w_m, \mathbf{m}_m, \mathbf{\Sigma}_m\}, m = 1, \ldots, M$, where we have mixture weight, mean vector and covariance matrix, respectively. We restrict the covariance matrices to be diagonal. The i-vector model is defined for the UBM component m as [27]:

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

where $\mathbf{V}_m$ is the sub-matrix of the total variability matrix, $\mathbf{y}$ is the latent vector, called an i-vector, ϵ_m is the residual term and $\mathbf{s}_m$ is the m'th sub-vector of the utterance dependent supervector. The ϵ_m is distributed as $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_m)$, where $\mathbf{\Sigma}_m$ is a diagonal matrix. Given all these definitions, posterior density of the $\mathbf{y}$, given the sequence of observed feature vectors, is Gaussian. Expectation of the posterior is the extracted i-vector. Hyperparameters of the i-vector model, $\mathbf{m}_m$ and $\mathbf{\Sigma}_m$, are copied directly from the UBM and $\mathbf{V}_m$ are estimated by the expectation maximization (EM) algorithm from the same corpus as is used to estimate the UBM.

The *cosine scoring* method is used to compare $\mathbf{w}_{\text{test}}$ and $\mathbf{w}_{\text{target}}$ i-vectors [27]. Cosine score of two i-vectors $\mathbf{w}_{\text{test}}$ and $\mathbf{w}_{\text{target}}$ is computed as their inner product $\langle \mathbf{w}_{\text{test}}, \mathbf{w}_{\text{target}} \rangle$, as

$$\mathrm{s}(\mathbf{w}_{\text{test}}, \mathbf{w}_{\text{target}}) = \frac{\hat{\mathbf{w}}_{\text{test}}^{\mathrm{T}} \ \hat{\mathbf{w}}_{\text{target}}}{\|\hat{\mathbf{w}}_{\text{test}}\| \ \|\hat{\mathbf{w}}_{\text{target}}\|}, \tag{2}$$

where $\hat{\mathbf{w}}_{\text{test}}$ is

$$\hat{\mathbf{w}}_{\text{test}} = \mathbf{A}^{\mathrm{T}} \mathbf{w}_{\text{test}}, \tag{3}$$

and $\mathbf{A}$ is the *heteroscedastic linear discriminant analysis* (HLDA) projection matrix [28] estimated from all training utterances. Further, $\hat{\mathbf{w}}_{\text{target}}$ is defined for a given dialect as,

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

where N_d is the number of training utterances in dialect d and $\hat{\mathbf{w}}_{id}$ is the projected i-vector of training utterance i from dialect d, computed the same way as (3). Obtaining $\{s_d, d = 1, \ldots, N\}$ scores for test utterances of dialect d, and total number of targeted models, N, scores are post-processed as [29]:

$$s'(d) = \log \frac{\exp(s_d)}{\frac{1}{N-1} \sum_{k \neq d} \exp(s_k)} \tag{5}$$

$s'(d)$ is the detection log-likelihood ratio and is used in the detection task.

3. Evaluation setup

3.1. Corpora

SAPU (Satakunta in Speech) corpus have been used to perform a series of experiments in this study. The data recorded in Satakunta, in southwestern Finland 2007-2013, in an interview setting. The topics were related to informants life and home region. Currently, the corpus consists of 282 recordings (231 hours 31 minutes)[1].

Satakunta region is divided into two distinctive dialectal regions, Southwestern dialects and the dialects of Häme. For our purposes, we selected 8 dialects — Luvia, Kokemäki, Honkilahti, Pori, Eurajoki, Rauma, Harjavalta, and Ulvila — with enough available data. All the audio files were partitioned into wave files of 30 seconds in duration, and downsampled to 8 kHz sampling rate. Table 1 shows the train and test files distributions within each dialects. There is no speaker overlap between training and test files.

[1]Corpus is located at the University of Turku the Syntax Archives server and is available by request.

Table 1: Training and test files distribution in the SAPU corpus.

Dialect	#Train files	#Test files	#Speakers
Luvia	386	315	31
Kokemäki	689	438	27
Honkilahti	845	413	24
Pori	341	289	15
Eurajoki	256	237	13
Rauma	237	64	9
Harjavalta	66	65	4
Ulvila	113	36	4

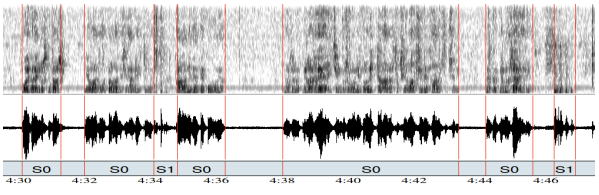

Figure 1: Speaker diarization scheme. S0 is the majority class (interviewee) and S1 is the minority class (interviewer).

3.2. Diarization

Two speakers are generally involved in the interviews. Speaker diarization [30] aims at 1) segmenting the audio stream into speech utterances, and 2) grouping all utterances belonging to the same speaker. In this study, diarization is mainly inspired by [31]. After noise reduction[2], a bidirectional audio source segmentation is applied using both *generalized likelihood ratio* (GLR) [32] and *Bayesian information criterion* (BIC). The resulting segments serve as an initial set of clusters that feed the clustering process.

This clustering is a variant of the Gaussian Mixture Model (GMM) system widely used for speaker recognition. The universal background model (UBM) is trained using all speech utterances from all recordings. To cope with short duration clusters, only 32 Gaussian components are used. Finally, the major cluster is selected and used for dialect recognition.

Fig. 1 shows the diarization scheme for a sample audio file. For this example, S0 is the majority class (interviewee) and S1 is the minority class (interviewer).

3.3. Measurement metrics

System performance is reported in terms of *equal error rate* (EER) and average detection cost (C_{avg}). EER corresponds to the operating point where false alarm and miss probabilities are equal. We report averaged EER across dialect-specific EERs. C_{avg} is defined as,

$$C_{\text{avg}} = \frac{1}{J}\sum_{j=1}^{M} C_{\text{DET}}(L_j), \tag{6}$$

where $C_{\text{DET}}(L_j)$ is the detection cost for subset of test segments trials for which the target dialect is L_j and J is the number of target languages. The per target dialect cost is computed as,

$$\begin{aligned} C_{\text{DET}}(L_j) &= C_{\text{miss}} P_{\text{tar}} P_{\text{miss}}(L_j) \\ &+ C_{\text{fa}}(1-P_{\text{tar}})\frac{1}{J-1}\sum_{k\neq j} P_{\text{fa}}(L_j, L_k) \end{aligned} \tag{7}$$

The miss probability (or false rejection rate) is denoted by P_{miss}, i.e., a test segment of dialect L_i is rejected as being in that dialect. On the other hand $P_{\text{fa}}(L_i, L_k)$ denotes the probability when a test segment of dialect L_k is accepted as being in dialect L_i. It is computed for each target/non-target dialect pairs. 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 dialect and was set to 0.5.

4. Results

4.1. Finnish dialect detection

We introduce speech attribute based systems in dialect recognition task and contrast it with baseline shifted delta cepstra and Mel frequency cepstral coefficients (SDC+MFCC), and single attribute (manner or place) system in Table 2. The parameters and combination (SDC and MFCC) were optimised in [6]. We also present results for attributes stacked across multiple frames. That is, we stack the estimated attribute feature vectors (either place or manner) across K neighboring frames to create a high-dimensional context feature vector. As discussed in detail in [7], the dimensionality of the context vector is reduced with principal component analysis (PCA). The PCA bases are trained from the same utterances as the universal background model (UBM), with 99% variance retained by the leading eigenvectors. In this work, we found that the PCA of context size $C = 10$ gives the best result on attributes. The PCA manner outperforms the baseline SDC+MFCC by 25% relative improvement considering C_{avg}. It also outperforms single manner and place attributes by 15% and 23% relative improvements, respectively. The place PCA is found not to be effective. This seems to contradict our earlier finding on another corpus [7]. While the exact reason is presently unknown, we note that the automatically determined PCA dimensionality for place attributes is smaller than in [7].

Literature of regional automatic dialect recognition is limited. In a study by DeMarco and Cox [15], SDC based i-vector system was used to classify fourteen British accents resulting 32% Id_{err}, which is comparable to 36% Id_{err} in Table 2. Later they improved the error rate to 19% by a very large scale fusion [16].

Table 2: Summary of results and compared against baseline spectral system, results are shown in average EER (Avg EER), C_{avg} and identification error rate (Id_{err}). C and d are context size and feature dimensionality, respectively.

Features (dimensionality)	Avg EER (%)	$C_{\text{avg}} \times 100$	Id_{err} (%)
SDC+MFCC (56)	14.20	5.31	36.08
Manner (18)	13.47	4.76	29.88
Place (27)	16.12	5.18	34.16
Manner+Place (45)	13.67	4.58	29.16
PCA Manner (C=10,d=30)	**12.52**	**4.00**	**29.01**
PCA Place (C=10,d=13)	17.60	5.64	37.65

[2]http://www1.icsi.berkeley.edu/Speech/papers/qio/

4.2. Levelling analysis

Here, we will further analyze the averaged detection results in terms of age groups. Fig. 2 presents the results per age group; that is, we choose a subset of original trials constrained to a given age group. We notice that the dialect in the younger age groups is considerably more difficult recognize than in the older age groups. The result indicates that the dialect of younger speakers has levelled. On the other hand, PCA manner considerably outperforms baseline SDC+MFCC for the youngest age group. It implies that attribute system is robust against the age related levelling for younger speakers.

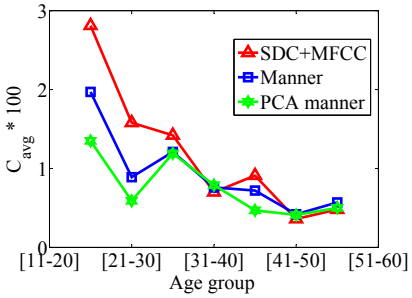

Figure 2: C_{avg} at different age groups.

We investigated more closely the cuts in the age group 11-20 that are correctly recognized by PCA Manner but incorrectly recognized by the spectral system, totalling 83 cuts from 19 different individuals. We show the example in Fig. 3, of one 30 seconds cut from a female speaker who is from Honkilahti municipality, however, in this cut she is recognized as being from Rauma by the spectral system. In this example, she says "*mum mielest se*" (in my opinion), where we notice word-final /n/ assimilated to bilabial nasal /m/. This would not happen in the Pori region dialects. Such an assimilation is typical for all the Southwestern Dialects (including Luvia and when preceded by bilabial phoneme /m/ Honkilanti). Of three detector scores per attribute we show here only the target score for clarity. We notice the nasal component is strong in the middle /m/, where dialectal difference shows.

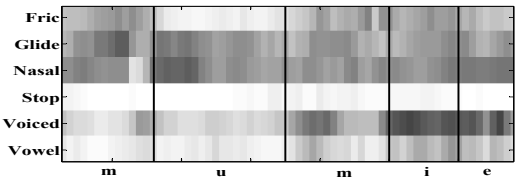

Figure 3: Target detection scores for the Manner of articulation detectors shown for the portion of "mum mielest se" (in my opinion).

It is interesting to see how much the attribute detection errors affect the dialect recognition performance for age group between 11 to 20 years old. Table 3 shows the confusion matrix of PCA manner system for this age group. Honkilahti is often misclassified as being Kokemäki, Pori, Eura; and Kokemäki being often misclassified as Ulvila. For Honkilahti dialects, the misclassification comes from the common prosodic features. on the other hand, Ulvila and Kokemäki are both Häme dialects.

Fig.4 shows how C_{avg} is affected by gender and region of birth for different systems. The dialectal differences of females is easier to recognize than for males. Similar to age analysis, PCA manner outperforms baseline SDC+MFCC and manner system. According to [33], various phonological and lexical forms and the syntactic-pragmatic features identified occur more often in women's than men's speech. Taking region of birth, results disagree with the common notion that those living in their home region have stronger dialects than those who have migrated from their home region. According to [34], language use of some migrated speakers show great situational variation. While there are always significant differences between the speakers of the same community, sometimes migrated speakers may speak even more dialectically. This kind of dialectal boosting appears specially in emphatic and affective occasions, when speaker talks with another person from the same region about the home region and people living there. The recordings of this corpus were recorded by the assistants born and raised in the same region.

Table 3: Confusion matrix of PCA manner system for age group between 11 and 20 years old. (There are no Eue, Rau and Har test utterances available for this age group.)

		Predicted label							
		Luv	Kok	Hon	Por	Eur	Rau	Har	Ulv
True label	Luv	35	3	7	4	6	10	1	1
	Kok	21	30	18	23	16	2	23	27
	Hon	23	41	168	48	57	23	24	20
	Por	7	0	8	23	8	0	1	6
	Ulv	2	7	7	3	0	1	1	9

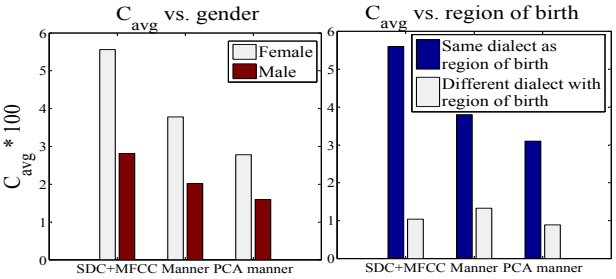

Figure 4: C_{avg} per gender and region of birth.

5. Conclusion

In this paper, we experimented with regional dialect recognition task. In terms of absolute error rates, it was shown to be a difficult task. There are two major sources of difficulty, differences between regional dialects are very small and the dialects are affected by the levelling phenomenon. Three levelling effects, age, gender and region of birth were studied in this paper. We showed that manner of articulation based recognition system can efficiently compensate the age levelling effect in Finnish dialect recognition. Furthermore, adding context information to manner attributes considerably improved the results.

6. Acknowledgements

This work was partly supported by Academy of Finland (projects 253000, 253120 and 283256) and Kone foundation.

7. References

[1] D. Britain, *Geolinguistics and linguistic diffusion*. Sociolinguistics: International Handbook of the Science of Language and Society, 2005.

[2] P. Kerswill and A. Williams, "Mobility and social class in dialect levelling: evidence from new and old towns in England," in *Dialect and migration in a changing Europe.*, Peter Lang, Frankfurt, 2000, pp. 1–13.

[3] E. Torgersen and P. Kerswill, "Internal and external motivation in phonetic change: Dialect levelling outcomes for an English vowel shift," *Journal of Sociolinguistics*, vol. 8, pp. 23–53, 2004.

[4] P. Kerswill, "Dialect levelling and geographical diffusion in British English," in *Social dialectology: in honour of Peter Trudgill*. Amsterdam: Benjamins, 2003, pp. 223–243.

[5] P. Kerswill and A. Williams, "Dialect levelling: change and continuity in Milton Keynes, Reading and Hull." in *Urban voices. Accent studies in the British Isles*, Arnold, London, 1999, pp. 141–162.

[6] H. Behravan, V. Hautamäki, and T. Kinnunen, "Foreign accent detection from spoken Finnish using i-vectors," in *INTERSPEECH*, Lyon, France, August 2013.

[7] H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee, "Introducing attribute features to foreign accent recognition," in *Acoustics, Speech and Signal Processing, 2014. ICASSP 2014. IEEE International Conference on*. IEEE, 2014.

[8] O. Scharenborg, M. J. Witteman, and A. Weber, "Computational modelling of the recognition of foreign-accented speech," in *INTERSPEECH*. ISCA, 2012.

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

[10] N. F. Chen, W. Shen, and J. P. Campbell, "A linguistically-informative approach to dialect recognition using dialect-discriminating context-dependent phonetic models," in *Acoustics, Speech and Signal Processing, 2014. ICASSP 2014. IEEE International Conference on*, 2010, pp. 5014–5017.

[11] F. Biadsy, J. Hirschberg, and M. Collins, "Dialect recognition using a phone-GMM-supervector-based SVM kernel." in *INTERSPEECH*. ISCA, 2010, pp. 753–756.

[12] N. F. Chen, W. Shen, J. P. Campbell, and P. A. Torres-Carrasquillo, "Informative dialect recognition using context-dependent pronunciation modelling." in *Acoustics, Speech and Signal Processing, 2011. ICASSP 2011. IEEE International Conference on*. IEEE, 2011, pp. 4396–4399.

[13] F. Biadsy, H. Soltau, L. Mangu, J. Navratil, and J. Hirschberg, "Discriminative phonotactics for dialect recognition using context-dependent phone classifiers," in *Odyssey*, 2010, p. 44.

[14] S. Sinha, A. Jain, and S. S. Agrawal, "Speech processing for Hindi dialect recognition," *Advances in Intelligent Systems and Computing*, vol. 264, pp. 161–169, 2014.

[15] A. DeMarco and S. J. Cox, "Iterative classification of regional British accents in i-vector space," in *MLSLP*, 2012, pp. 1–4.

[16] A. DeMarco and S. J. Cox., "Native accent classification via i-vectors and speaker compensation fusion," in *INTERSPEECH*, 2013, pp. 1472–1476.

[17] 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.

[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.-Y. Chiang, S. M. Siniscalchi, Y.-R. Wang, S.-H. Chen, and C.-H. Lee, "A study on cross-language knowledge integration in Mandarin LVCSR," in *Proc. ISCSLP*, HONG KONG, Dec. 2012, pp. 315–319.

[21] C.-Y. Chiang, S. M. Siniscalchi, , S.-H. Chen, and C.-H. Lee, "Knowledge integration for improving performance in LVCSR," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 1786–1790.

[22] 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.

[23] 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.

[24] 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.

[25] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Toward a detector-based universal phone recognizer," in *Proc. ICASSP*, Las Vegas, NV, USA, Mar./Apr. 2008, pp. 4261–4264.

[26] P. Schwarz, P. Matějaka, and J. Cernock, "Hierarchical structures of neural networks for phoneme recognition," in *ICASSP*, Toulouse, France, 2006, pp. 325–328.

[27] N. Dehak, P. 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.

[28] M. Loog and R. P. W. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: The chernoff criterion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 732–739, 2004.

[29] N. Brummer and D. Van Leeuwen, "On calibration of language recognition scores," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, June 2006, pp. 1–8.

[30] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 356–370, Feb 2012.

[31] E. Khoury, C. Senac, and J. Pinquier, "Improved speaker diarization system for meetings," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4097–4100.

[32] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Multistage speaker diarization of broadcast news," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1505–1512, Sept 2006.

[33] R. Lakoff, *Language and woman's place*, ser. Harper colophon books. Harper & Row, 1975.

[34] P. Nuolijrvi, *Kieliyhteisn vaihto ja muuttajan identiteetti.*, ser. Tietolipas. Helsinki: SKS., 1986.