2015/2500M-0036

ISSN 1797-3457 (verkkojulkaisu)
ISBN 978-951-25-2758-8 (PDF)

SUMMARY REPORT

Foreign accent recognition project (FARP)

Ville Hautamäki,
University of Eastern Finland (UEF), School of Computing
villeh@cs.uef.fi

Abstract

If a traveler or immigrant has biometric passport, then verifying the validity of the passport can be done by measuring face, fingerprint or voice and matching it against the stored model in the chip in the passport. However, many countries do not have not implemented yet biometric passport technology, in addition some Finnish passport categories, such as temporary passport do not have the chip. In this project we study a form a form of soft-biometric, which does not identify an individual but a group of individuals. In this work we concentrated on the country origin, that is a required field in all passports. If we detect the country of origin to be false then we say that possibly the passport is fake. We implement the country of origin detection by the way of detecting the English foreign accent. We have recently proposed a universal acoustic characterisation to foreign accent recognition, in which any spoken foreign accent was described in terms of a common set of fundamental speech attributes. Although experimental evidence demonstrated the feasibility of our approach, we believe that speech attributes, namely manner and place of articulation, can be better modelled by a deep neural network. In this work, we propose the use of deep neural network trained on telephone bandwidth material from different languages to improve the proposed universal acoustic characterisation.

1. Introduction

If a traveler or immigrant has biometric passport, then verifying the validity of the passport can be done by measuring face, fingerprint or voice and matching it against the stored model in the chip in the passport. However, many countries do not have not implemented yet biometric passport technology, in addition some Finnish passport categories, such as temporary passport do not have the chip. In this project we study a form a form of soft-biometric, which does not identify an individual but a group of individuals. In this work we concentrated on the country origin, that is a required field in all passports. If we detect the country of origin to be false then we say that possibly the passport is fake. One famous case of fake passport travelers were two Iranian individuals in Malaysia Airlines flight MH 370, who were traveling with stolen passports. On of them was using Australian and the other Italian passport. In that case the optimal response of our proposed system would have been that it does not seem that this traveler is from Italy.

We implement the country of origin detection by the way of detecting the English foreign accent. Our claim is that the utilization of speech attributes, namely manner and place of articulation, will not only help in detection of the accent but also reveal the way how some ethnic group typically mispronounces some English words. To exemplify, Italians often do not

aspirate the /h/ sound in words such as *house*, *hill* and *hotel*. This lack of aspiration works as cue that we have person originating from Italy speaking in English. However, estimation of the manner and place of articulation is difficult task, where some attributes are not very well detected at all. So the main task taken in this project was to significantly improve on the speech attribute detection accuracy.

2. Research objectives and accomplishment plan

In automatic foreign accent recognition the mother tongue (L1) of non-native speakers has to be recognised given a spoken segment in a second language (L2) [1]. We may think of L1 recognition as a language recognition task [2], where L1 is the target language to be recognised. However, language recognition techniques based on n-gram phoneme statistics are not directly usable, as the collected phoneme statistics would match the L2 language. In [3], it was advocated the use of speech attributes, namely manner and place of articulation, to universally characterise all language and accents, and experimental evidence proved their effectiveness in foreign accent recognition. Foreign accent variation is a nuisance factor that negatively affects automatic speech, speaker and language recognition systems [4,5]. Most of the speech technology systems have been tailored to native speech, but those systems rarely work well on non-native or accented speech, such as the *automatic speech recognition* (ASR) [6,7].

The most direct way to overcome the problem of non-native speech is to train separate statistical models for each L1-L2 pair. But by using the accent universal units, we would be able to *compensate* against the L1 nuisance effects. Similarly, such units can be used in foreign accent conversion [8] with the idea of reducing the perceptual effect of accentedness. In [8], the accent universal units were articulatory gestures, namely manner and place of articulation recorded using the *electromagnetic articulography* (EMA). Accent conversion is achieved by obtaining parallel audio and EMA recordings from the L1 and L2 targets. Being limited to EMA recordings to obtain articulatory gesture scores is by its vary nature practically very restricted. The *automatic speech attribute transcription* (ASAT) framework [9], is bottom-up detection-based framework, where speech attributes are extracted using data-driven machine. We were able to successfully use these detector scores in foreign accent recognition [3], and regional dialect recognition [10] by modeling the stream of detector scores using the i-Vector methodology [11]. In contrast to phonotactic language recognition systems, the i-Vector based method defers all decisions until the final accent recognition is made. Experimental results demonstrated the effectiveness of our i-Vector modelling of attributes, and a significant system performance improvement over conventional spectrum-based techniques was demonstrated on the Finnish national foreign language certificate corpus. Nonetheless, we also observed that some speech attributes were not properly modelled by the sh*allow neural networks* (SNN), employing a single-hidden non-linear layer. In fact, the baseline speech attribute front-end exhibit a large error rate variance [12].

We believe that accent recognition accuracy can be greatly enhanced if more powerful data-driven learning systems replace shallow networks for speech attribute modelling. *Deep neural networks* (DNNs), e.g., [13], have been successfully applied across a range of different speech processing tasks in recent years, such as conversational-style speech recognition, e.g., [14], noise robust applications [15], multi- and cross-lingual learning techniques, e.g., [16]. Inspired by the success of those applications, we want to explore the use of DNNs to extract manner and place of articulation attributes to be used in automatic accent recognition systems. DNNs are chosen because they (i) can be easily trained on high dimensional features, (ii) have the potential to learn more efficient and effectively non-linear feature mappings, and (iii) may better capture the complex relationships among speech

attributes. Two speech attribute classifiers for manner and place of articulation, respectively, are built using DNNs trained on telephone bandwidth speech material from the six different languages in the OGI Multi-language Telephone Speech corpus [17].

3. Materials and methods

Manner of articulation classes, namely, glide, fricative, nasal, stop, and vowel, and *place of articulation* classes, namely coronal, dental, glottal, high, labial, low, mid, palatal, and velar, are the speech attributes used in this work. Speech attributes can be obtained for a particular language and shared across many different languages, and they can thereby be used to derive a universal set of speech units. Furthermore, data-sharing across languages at the attribute level is naturally facilitated by the nature of these classes as shown in [18]. In [19], the authors have demonstrated that manner and place of articulation attributes can compactly characterise any spoken language along the same lines as in the ASAT paradigm for ASR [9]. Furthermore, it was shown that off-the-shelf data-driven attribute detectors built to address automatic language identification tasks [18] can be employed without either acoustic adaptation or re-training for characterising speaker accents never observed before [3]. In [3], attribute detectors were built using shallow neural networks, namely single-hidden layer, feed-forward neural networks. Here we want to test deeper architectures.

Inspired by the success of those applications, here we want to explore the use of DNNs to extract manner and place of articulation attributes to be used in automatic accent recognition systems. DNNs are chosen because (i) can be easily trained on high dimensional features, (ii) have the potential to learn more efficient and effectively non-linear feature mappings, and (iii) may better capture the complex relationships among speech attributes. Using the DNNs we can estimate posterior probability of a speech attribute per each frame (around 20 ms cut of the audio). The outcome of this processing is variable length stream of vectors from probability simplex. One scalar represents the posterior probability of one speech attribute. To be able to classify each utterance in predefined foreign accent classes we still need to extract a fixed length vector representation. This is called an i-vector.

The idea behind i-vector model is that the feature vectors x_i, $i=1,...,N$, where N is the number of speech attribute feature vectors, can be compressed into a fixed length vector. All variability, such as accent speaker and channel, are retained in that representation of an utterance. For that reason, i-Vector model is also called total variability modeling [11]. It stems from the idea that feature stream can be modeled by *Gaussian mixture model* (GMM) that is adapted by relev*ance maximum a posteriori* (MAP) from the *universal background model* (UBM). Then stacking the adapted GMM mean vectors creates a fixed length representation of the utterance. But the dimensionality of the GMM supervector space is very high, easily more than 100000. In the i-Vector model, the utterance dependent supervector is defined as s = m + Tw + noise, where m is the utterance independent mean vector, copied from the UBM by stacking the mean vectors, T is a rectangular low rank matrix and the latent vector w is distributed according to N(0,I), the T represents the captured variabilities in the supervector space and noise captures the residual variability. The residual is distributed N(0,$\mathbf{\Sigma}$), where $\mathbf{\Sigma}$ is copied directly from the GMM. The T-matrix is estimated from the held-out corpus, typically same as the where UBM is estimated from, via an *expectation maximization* (EM) algorithm. The idea of the algorithm is that we infer w, which is the posterior mean, for each training utterance given an estimate of T-matrix and then estimate new T-matrix and so on. The estimation is very CPU intensive, so typically only few, for example five, iterations is used in practice. We use *cosine scoring* to measure similarity of two i-Vectors [11]. Target accent model is nothing else than average i-vector of all training set utterances for a given accent.

4. Results and discussion

The front-end is built using two independent DNNs having six hidden layers and 1024 hidden nodes. The input feature vector is a 45-dimension mean-normalized log-filter bank feature with up to second-order derivatives and a context window of 11 frames, forming a vector of 495-dimension 45 x 11 input. The number of output classes is equal to 6 for manner, and 10 for place. In addition, a further output class is added to both DNNs to handle possible unlabelled frames. The DNN was trained with an initial learning rate of 0.008 using the cross-entropy objective function. It was initialised with the stacked *restricted Boltzmann machines* (RBM) by using layer by layer generative pre-training. An initial learning rate of 0.01 was then used to train the Gaussian-Bernoulli RBM and a learning rate of 0.4 was applied to the Bernoulli-Bernoulli RBMs. This DNN architecture follows conventional configurations used in the speech community, and it was not optimised for the corpora and task at hand. The ``stories'' part of the OGI Multi-language telephone speech corpus [17] was used to train the attribute 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 of training and 0.52 hours of validation data.

In Table 1, we report manner and place of articulation accuracies for each specific attribute using either one or six hidden layers. Classification accuracies increased consistently for all attributes except silence when moving from one to six hidden layers, as we expected. It should be noted that although silence classification accuracy does not increase, it is already above 90%. Glide and dental are instead still very hard to detect, and even with 6 hidden layers an accuracy of only 30% can be attained.

Table 1: Detection results for Manner (left) and Place (right). Contrasting shallow network with 1 layer versus deep network with 6 layers.

Attribute	1 layer	6 layers
fricative	69.2	**72.0**
glide	27.6	**30.0**
nasal	75.3	**76.8**
silence	**92.5**	92.3
stop	72.3	**75.4**
vowel	91.4	**91.6**
Total	79.2	**80.1**

Attribute	1 layer	6 layers
coronal	55.0	**57.7**
dental	27.7	**32.5**
glottal	39.1	**43.3**
high	54.0	**56.5**
labial	53.3	**56.4**
low	66.0	**68.5**
mid	61.6	**62.3**
palatal	42.3	**45.6**
silence	**93.8**	93.4
velar	49.4	**56.2**
Total	61.8	**63.7**

To better appreciate experimental results reported in this paper, we compared our attribute-based systems against two spectral-based accent recognition systems based on SDC and MFCC feature vectors, respectively, which have proven to give best performance in foreign accent recognition tasks [20]. Accent classifiers in these two systems were built using either GMM-UBM [21] or i-Vector approach. According to [20], the UBM size was set to 512, i-

Vector dimension to 1000 and HLDA output dimension to 180. The UBM and T-matrix were estimated from the same held-out set, not used in either training or testing the foreign accent models. The experiments are performed using a subset of NIST SRE 2008 corpus. In our experiments, we selected the test utterances from the original 10sec NIST SRE 2008 cuts in order to keep the test setup in line with the standard language and accent recognition test.

Table 2: Average equal error rate (EER) results on foreign accent detection task.

Feature (dimension)	Classifier	$EER_{avg}(\%)$	$C_{avg} \times 100$
SDC+MFCC (56)	GMM-UBM	16.94	9.00
SDC+MFCC (56)	i-Vector	13.82	7.87
SNN Place (27)	i-Vector	12.00	7.27
DNN Place (11)	i-Vector	11.11	7.00
SNN Manner (18)	i-Vector	11.09	6.70
DNN Manner (7)	i-Vector	**10.45**	**6.50**

Above results indicate the effectiveness of the DNN attribute features over spectral SDC+MFCC and attribute features. Next, we compare the language-wise results achieved by shallow and deep architecture for the manner case for the both the Finnish and the English task. We compensate against the lack of data, by performing a jack-knifing type evaluation. More details of the experimental setup can be found in [22]. Table 3 shows per-accent recognition accuracy on the English task. In both Manner and DNN manner systems, Cantonese attains the lowest recognition accuracy with EER of 17.68% and 13.50%, respectively; and the easiest accent is Thai with EER of 11.70% and 9.31%, respectively, in both systems.

Table 3: Showing detection results per target foreign accent in terms of equal error rate (EER). Results are obtained via jack-knifing to get reliable per target accent error rates.

Accent	SNN Manner	DNN Manner
Cantonese	17.68	13.50
Hindi	15.75	13.15
Vietnamese	15.44	12.22
Russian	13.16	10.00
Korean	12.54	11.97
Japanese	11.75	10.76
Thai	11.70	9.31
Total (average)	14.00	11.55

Results show clear improvement in the per language detection accuracy, when more accurate speech attribute estimation methods are used. Overall, we are able to obtain around 10% equal error rate level in foreign accent detection. We are confident that application of the *recurrent neural network* (RNN) instead of the i-vector method we would be able to significantly improve on these current results. The variants of the RNN modeling have been found this year to provide state-of-the-art performance in automatic speech recognition. It is able to model long-term frame dependencies that are very difficult for GMM based models, such as HMM and i-vector models. In addition, non-linearity can be increased quite easily as in other deep learning models. In addition to obtaining more accurate model for speech attribute feature streams, we are currently pursuing better ways to estimate

speech attributes themselves. One fairly clear idea is to estimate manner and place jointly so that detection of manner will help with place and vice versa. These topics are currently under investigation by our team and hopefully we can report new results early next year.

We have to note that results obtained here are from English telephone conversational speech. In the deployed system we would use wideband recorded audio, where target foreign accent models have been estimated from the actual material we will observe in the deployed conditions. Currently, we are developing a demonstrator app where self collected speech data will be used. We will release the app when reasonable amount of data has been collected so that the recognition system is accurate enough.

5. Conclusions

We managed to improve the foreign accent recognition performance on the English telephone speech. However, the current recognition error rates still hovers at around 10%, which means that in the border control use case 10% of the false passport detections would actually be false alarms. Extra workload given to border control officer thus would be too much as a lot of extra effort would be needed to verify all false alarms also. So, in order to apply our techniques in the MATINE case, we would need to improve the recognition accuracy and find out minimum error rate users would be satisfied with our system. In order to improve recognition accuracy we could consider also other soft biometrics that can be recognized from the speech signal, such as age and gender. Final system would then be a fusion of all these separate recognizers.

6. Scientific publishing and other reports produced by the research project

The complete foreign accent system with speech attributes is presented in this work. Speech attribute detection is shallow neural network.

Hamid Behravan, Ville Hautamäki, Sabato Siniscalchi, Tomi Kinnunen, and Chin-Hui Lee, "i-Vector Modeling of Speech Attributes for Automatic Foreign Accent Recognition", *IEEE Transactions on Audio, Speech and Language Processing,* 2016 (accepted).

In this work, we improved the speech attribute detection by the way of using modern deep neural network stragies. This resulted in improvement in foreign accent recognition performance.

Ville Hautamäki, Sabato Siniscalchi, Hamid Behravan, Valerio Mario Salerno and Ivan Kukanov, "Boosting Universal Speech Attributes Classification with Deep Neural Network for Foreign Accent Characterization", Interspeech 2015, Dresden, Germany, September 2015 (accepted).

7. References

[1] H. Hansen and L. M. Arslan, “Foreign accent classification using source generator based prosodic features,” in Proc. of ICASSP, 1995, pp. 836–839.
[2] 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, 2013.

[3] H. Behravan, V. Hautam¨ki, 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.
[4] L. M. Arslan and J. H. Hansen, “Language accent classification in American English,” Speech Communication, vol. 18, no. 4, pp. 353–367, 1996.
[5] P. Angkititraku and J. H. Hansen, “Advances in phone-based modeling for automatic accent classification,” in IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 2, 2006, pp. 634–646.
[6] V. Gupta and P. Mermelstein, “Effect of speaker accent on the performance of a speaker-independent, isolated word recognizer,” J. Acoust. Soc. Amer., vol. 71, no. 1, pp. 1581–1587, 1982.
[7] R. Goronzy, S. Rapp, and R. Kompe, “Generating non-native pronunciation variants for lexicon adaptation,” Speech Communication, vol. 42, no. 1, pp. 109–123, 2004.
[8] S. Aryal and R. Gutierrez-Osuna, “Accent conversion through cross-speaker articulatory synthesis,” JASA, vol. 137, no. 1, pp. 433–446, January 2015.
[9] 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.
[10] H. Behravan, V. Hautam¨ki, S. M. Siniscalchi, E. el Khoury, T. Kurki, T. Kinnunen, and C.H. Lee, “Dialect levelling in finnish: a universal speech attribute approach,” in INTERSPEECH 2014, 2014, pp. 2165–2169.
[11] 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.
[12] V. H. Do, X. Xiao, V. Hautamäki, and E. S. Chng, “Speech attribute recognition using context-dependent modeling,” in APSIPA ASC, Xi’an, China, October 2011.
[13] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.
[14] F. Seide, G. Li, and D. Yu, “Conversational speech transcription using context-dependent deep neural networks.” in Proc. Interspeech, 2011, pp. 437–440.
[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” ACM/IEEE Trans. Audio Speech and Lang. Processing, vol. 23, no. 1, pp. 7–19, 2014.
[16] Y. Miao and F. Metze, “Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training.” in Proc. Interspeech, 2013, pp. 2237–2241.
[17] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, “The ogi multi-language telephone speech corpus,” in Proc. of ICSLP’92, 1992.
[18] 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.
[19] 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.
[20] H. Behravan, V. Hautamäki, and T. Kinnunen, “Foreign accent detection from spoken Finnish using i-vectors,” in Proc. of INTERSPEECH, 2013, pp. 79–83.
[21] P. Torres-Carrasquillo, T. Gleason, and D. Reynolds, “Dialect identification using Gaussian mixture models,” in Proc. of Odyssey, 2004, pp. 757–760.
[22] H. Behravan, V. Hautamäki, and T. Kinnunen, “Factors affecting i-vector based foreign accent recognition: a case study in spoken Finnish,” Speech Communication, vol. 66, pp. 118–129, 2015.