Contrasting the Effects of Different Frequency Bands on Speaker and Accent Identification

Saeid Safavi, Abualsoud Hanani, Martin Russell, *Member, IEEE*, Peter Jančovič, *Member, IEEE*, and Michael J. Carey

***Abstract*—This letter presents an experimental study investigating the effect of frequency sub-bands on regional accent identification (AID) and speaker identification (SID) performance on the ABI-1 corpus. The AID and SID systems are based on Gaussian mixture modeling. The SID experiments show up to 100% accuracy when using the full 11.025 kHz bandwidth. The best AID performance of 60.34% is obtained when using band-pass filtered (0.23–3.4 kHz) speech. The experiments using isolated narrow sub-bands show that the regions (0–0.77 kHz) and (3.40–11.02 kHz) are the most useful for SID, while those in the region (0.34–3.44 kHz) are best for AID. AID experiments are also performed with intersession variability compensation, which provides the biggest performance gain in the (2.23–5.25 kHz) region.**

***Index Terms*—ABI-1, accent identification, bandwidth, frequency region, Gaussian mixture model, intersession variability compensation, speaker identification, sub-band.**

I. Introduction

It is well known that an acoustic speech signal contains information beyond its linguistic content. This paralinguistic information, includes clues to a speaker's accent and identity, which are exploited by automatic accent identification (AID) and speaker identification (SID) systems. The relationship between AID and SID is asymmetric, since accent information is relevant to SID but speaker information is a distraction in the context of AID.

Currently, the most commonly used parameterisation for both AID and SID is to represent a spoken utterance as a sequence of Mel-Frequency Cepstral Coefficient (MFCC) vectors, covering the entire frequency bandwidth. However, it is known that different frequency regions contain different types of information. For instance, the SID study in [1], which was performed on the clean TIMIT corpus using mono Gaussian modeling, showed that the frequency regions below 600 Hz and above 3000 Hz provided better SID accuracy than the middle-frequency regions. However no similar study has been conducted for AID. In this letter, the contrasting importances of different frequency bands for AID and SID are investigated, using contemporary GMM-based systems.

In the most widely used approach to AID and SID, the distributions of feature vectors are characterized using a Gaussian Mixture Model (GMM) [2], [3]. Individual accent or speaker GMMs are typically built by MAP adaptation of a Universal Background Model (UBM). This is an accent and speaker independent GMM constructed using data from a variety of accents, speakers and background conditions. This approach has been very effective for various SID tasks [4] and its performance remains comparable to that obtained with more complex models. It has also been applied to AID, but with less success [5]. An alternative is to use a discrimination-based approach, such as a Support Vector Machine (SVM), applied to GMM super-vectors, which consist of the 'stacked' means of the mixture components of the accent or speaker GMMs [5], [6]. A combination of the GMM and SVM approaches for SID, where the GMM was used to calculate likelihood values and the SVM classifier was used to separate the likelihood values for a correct speaker and impostor, is presented in [7] and showed slightly better results than the GMM system alone. Other acoustic-based approaches incorporate the use of phone durations and average cepstra [8], phone and word-level Hidden Markov Models (HMMs) [9]–[12], and stochastic trajectory models [13]. The most successful systems use intersession variability compensation (ISVC). This is a subspace projection technique for removing irrelevant variability in speech classification tasks, which has been shown to improve the performance of speaker, language and accent identification and has become a standard component of these systems [5], [17]. 'Phonotactic' approaches to AID exploit accent-dependent differences in the sequences in which speech sounds occur [14]. For AID, these approaches outperform the GMM-based acoustic methods described above [5].

This letter begins by reporting the results of applying GMM-based classifiers to AID (14 classes) and SID (93 or 94 classes) on the Accents of the British Isles (ABI) speech corpus [15]. SID accuracy using full bandwidth (11.025 kHz) speech and 512 component GMMs is close to 100%, confirming that SID is a relatively simple task for this type of data [1]. The best AID accuracy is 60.34%, obtained when the recordings are band-pass filtered (0.23–3.4 kHz), using 4096 component GMMs and a pitch-based speech activity detector (SAD).

Next, AID and SID experiments are conducted on the same narrow sub-bands. In this way it is possible to contrast the utility of information in different regions of the speech spectrum for these tasks. The results show that it is useful to divide the spectrum into four regions: A (0 to 0.77 kHz), B (0.34 to 3.44 kHz), C (2.23 to 5.25 kHz) and D (3.40 to 11.02 kHz). Our results suggest that speaker information dominates in regions A, corresponding to primary vocal tract resonance information, and D,

Manuscript received May 01, 2012; revised July 27, 2012; accepted September 14, 2012. Date of publication October 02, 2012; date of current version October 10, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ali Taylan Cemgil.

S. Safavi, M. Russell, P. Jančovič, and M. J. Carey are with the School of Electronic, Electrical, and Computer Engineering, The University of Birmingham, B15 2TT Birmingham, U.K. (e-mail: sxs796@bham.ac.uk; aah648@bham.ac.uk; m.j.russell@bham.ac.uk; p.jancovic@bham.ac.uk; m.carey@bham.ac.uk).

A. Hanani was with the School of Electronic, Electrical, and Computer Engineering, The University of Birmingham, B15 2TT Birmingham, U.K.. He is now with the Faculty of Information Technology, Birzeit University, West Bank, Palestine (e-mail: ahanani@birzeit.edu).

Digital Object Identifier 10.1109/LSP.2012.2221697

corresponding to high-frequency sounds. These results are consistent with [1]. In contrast, region B is most useful for AID, indicating that the vocal tract resonance information in this region is biased towards linguistic, rather than speaker information. Region C contains both types of information, although speaker information appears to dominate. When ISVC is applied to sub-band AID systems the biggest gain is observed in region C, where AID performance is improved by 24%, indicating that ISVC is able to factor out some of the speaker information in this region.

II. Corpus and Data Description

The "Accent of the British Isles" corpus of regionally accented adult's speech (ABI-1) was used in all experiments. This was collected to support research into the implications of regional accents for speech and language technology. ABI-1 [15] comprises recordings of speech representing thirteen regional accents of British English, namely Belfast, Birmingham, Burnley (Lancashire), Denbigh (North Wales), Elgin (Scottish Highlands), Dublin, Glasgow, Hull (East Yorkshire), Inner London, Liverpool, Lowestoft (East Anglia), Newcastle and Truro (Cornwall), plus 'Standard Southern English' (SSE). The objective in each location was to record twenty subjects (ten men and ten women) who were born in the location and had lived there for all of their lives. Each subject recorded approximately 15 minutes of read speech. The prompt texts were chosen for their relevance to applications or their phonetic content. The recordings were made in relatively quiet rooms. The ABI-1 corpus comprises recordings of 283 subjects, of which 280 were used in the current study. ABI-1 consists of approximately 70 hours of recordings, with speakers' ages ranging from 16 to 79 years. The recordings are transcribed at the phrase level, but the transcriptions were not used in the present study. All ABI-1 recordings were made using head mounted and desk microphones, and sampled at 22.05 kHz. Only the head mounted microphone recordings were used for experiments reported in this letter.

For both SID and AID, the speakers were divided into three subsets; two with 93 and one with 94 speakers. Gender and accent were distributed equally in each subset. A "jack-knife" training procedure was used in which two subsets were used for training and the remaining subset for testing. This procedure was repeated three times with different training and test sets, so that each ABI-1 speaker was used for testing, and no speaker appeared simultaneously in the training and test sets. The SID systems were evaluated using 993 segments of length 3, 10, and 30 seconds from all test recordings. The AID systems were evaluated using 1504 30-seconds segments from all test recordings. All the above indicated segment lengths are after silence removal.

III. System Description

A. Signal Analysis

Feature extraction for both SID and AID was performed as follows. Periods of silence were discarded using an energy-based SAD. The speech was then segmented into 20-ms frames (10-ms overlap) and a Hamming window was applied. The short-time magnitude spectrum, obtained by applying the FFT, is passed to a bank of 31 Mel-spaced triangular band-pass filters, spanning the frequency region from 0 Hz to 11025 Hz. The center frequency for each filter is shown in Table I.

TABLE I
The Center Frequencies for 31 Mel-spaced Band-Pass Filters

FILTER NUMBER	CENTER FREQUENCY (Hz)	FILTER NUMBER	CENTER FREQUENCY (Hz)
1	129	17	2239
2	258	18	2497
3	344	19	2799
4	473	20	3100
5	559	21	3445
6	645	22	3832
7	775	23	4263
8	861	24	4737
9	990	25	5254
10	1076	26	5857
11	1205	27	6503
12	1335	28	7235
13	1464	29	8053
14	1636	30	8957
15	1808	31	9948
16	2024		

The SID and AID experiments were first performed using the full bandwidth (0–11.025 kHz) and telephone bandwidth (0.23–3.4 kHz) speech. The latter was obtained by passing the recordings through a band pass filter. The calculation of MFCCs was based on all 31 filters and the first 23 filters for full and telephone bandwidth, respectively. In both cases, the first 19 MFCCs were used.

To investigate the effect of different frequency regions, separate SID and AID experiments were conducted using frequency band limited speech data comprising the outputs of groups of four adjacent filters. We considered 28 overlapping sub-bands, where the Nth sub-band comprises the outputs of filters N to $N+3$ ($N = 1, \ldots, 28$). Each set of four filter bank outputs was transformed to 4 Mel-frequency cepstral coefficients (MFCCs) and feature warping [16] was applied.

B. GMM-UBM System

The SID and AID systems are based on the Gaussian Mixture Model – Universal Background Model (GMM-UBM) method [3]. In what follows, "class" refers to accent or speaker, depending on the particular experiment.

In the GMM-UBM approach, a UBM is built using utterances from the training sets of all classes. Class-dependent models are obtained by MAP adaptation [3], adapting the means of the UBM, using the class-specific enrollment data. The result is one UBM and C class-dependent GMMs, where in our experiments C is 14 and 93 or 94 (depending on the "jack-knife" set) for AID and SID, respectively.

For AID, the inter-session variability within a class, such as inter-channel and inter-speaker variability, is estimated using the technique described in [17]. This technique is applied on a band specific level with a fixed band-independent rank. During testing, both the UBM and the class dependent GMM means are shifted in the estimated 'nuisance' direction of each test utterance before scoring. ISVC was not used in the full bandwidth AID experiments (Table II) or the initial sub-band experiments (Figs. 2 and 3) but was used in the final AID sub-band experiments (Figs. 4 and 5). ISVC was not used in the SID systems, because the same recording equipment was used for the whole corpus and the recording environment for each subject did not change.

TABLE II
SUMMARY OF RESULTS FOR SID AND AID SYSTEMS (IDENTIFICATION RATE)

GMM	AID (30 sec)		SID (30 sec)		SID (10 sec)		SID (3 sec)	
comp.	Full BW	Tel BW	Full BW	Tel BW	Full BW	Tel BW	Full BW	Tel BW
512	38.50	57.50	100	97.54	100	95.09	98.98	88.18
2048	40.64	59.42						
4096	42.54	60.34						

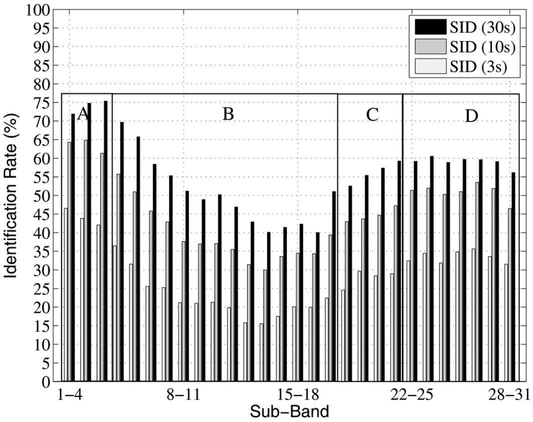

Fig. 1. SID performance as a function of frequency sub-band for 3, 10 and 30 second test signals when using 512 component GMMs.

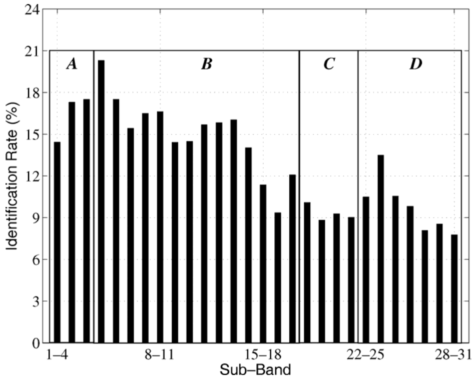

Fig. 2. AID performance as a function of frequency sub-bands using 512 component GMMs and 30 second test signals.

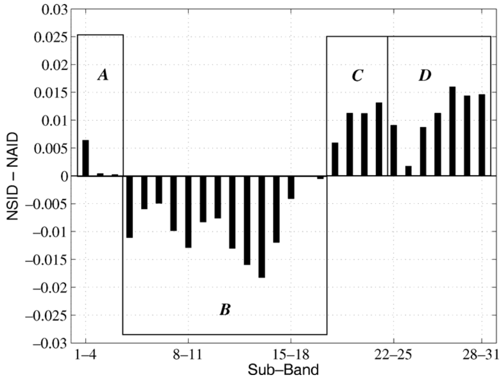

Fig. 3. The difference between the normalized SID and AID performance for frequency sub-bands using 30 second test signals.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In order to demonstrate the competitiveness of our AID and SID recognition systems, the first experiments are performed using the full bandwidth (0–11.025 kHz) and telephone bandwidth (0.23–3.4 kHz) speech. The obtained results are presented in Table II. The AID system here uses a pitch-based SAD. Using the full bandwidth speech, the performance of the AID system using 30 second test segments is 38.50%, 40.64% and 42.54% with 512, 2048 and 4096 mixture components, respectively. The performance of the SID system using a 512 component GMM is 98.98%, 100% and 100% for 3, 10 and 30 second test files, respectively. When using the simulated telephone bandwidth speech, the AID performance increased by between 42% and 49%, while the SID performance dropped by between 3% and 11%. Current state-of-the-art AID systems typically use 4096 component GMMs and pitch-based SAD which also achieved the best performance (60.34%) in our experiments (the corresponding performance for energy-based SAD is 57.37%).

In the following, we study the performance of the SID and AID systems for each individual sub-band. For the purpose of analysis, it is also useful to divide the entire frequency range into four broader regions: A from 0 to 0.77 kHz, B from 0.34 to 3.44 kHz, C from 2.23 to 5.25 kHz and D from 3.40 to 11.02 kHz. Both the AID and SID systems here use the energy-based SAD (as the pitch-based detector would eliminate most of the high-frequency unvoiced fricative sounds). Based on the results in Table II and since the feature dimensionality is much lower when using individual sub-bands as opposed to the full bandwidth, both the SID and AID systems in the following experiments are based on 512 component GMMs. The results for SID are shown in Fig. 1. It can be seen that the lowest performance is obtained using the mid frequency sub-bands (region B). These results are consistent with previous findings reported in [1], which were obtained for clean speech on the TIMIT corpus and using only mono Gaussian modeling. The performances for 3, 10 and 30 second test files show similar trends, but accuracy is around 27% and 10% lower on average for the 3 and 10 second test data, respectively, compared to the 30 second data. The results for AID are shown in Fig. 2. Compared with the results for SID, region B appears to be more useful, while regions C and D are less useful.

In order to contrast the SID and AID performances, the results presented in Figs. 1 and 2 were first normalized to sum to one over all the sub-bands and then subtracted. The resulting contrastive SID and AID performance, which we refer to as *NSID – NAID*, is shown for 30 second test data in Fig. 3. Regions with positive values (A, C and D) in Fig. 3 carry more speaker specific information than accent information, while the region with negative values (B) contains more AID information. Region A, corresponding to the primary vocal tract resonance information of vowel and nasal sounds, and regions C and D, corresponding to high frequency sounds such as fricatives, are most useful for SID. Region B is where one would expect to find vocal tract resonance information for general voiced speech sounds. Although this information will be biased by individual differences in vocal tract physiology, linguistic information dominates and makes the region most useful for AID. This is consistent with observations on the importance of vowels in subjective analyses of accent [18].

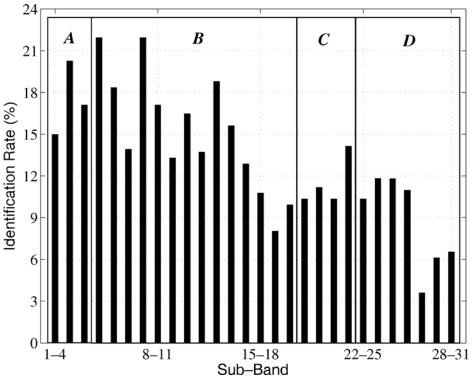

Fig. 4. AID performance after ISVC as a function of frequency sub-bands using 512 component GMMs and 30 second test signals.

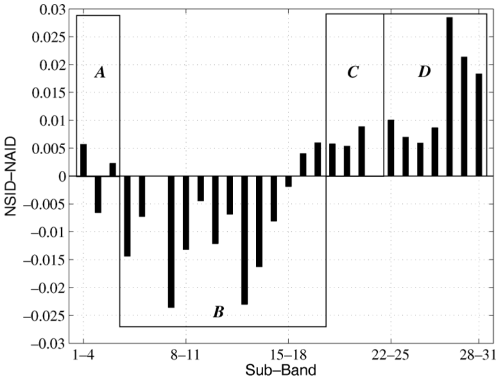

Fig. 5. The difference between the normalized SID and AID performance for frequency sub-bands (after application of ISVC to AID system).

Last, we present results obtained with ISVC [17]. To investigate which frequency bands gain most from ISVC in AID, we apply it to each frequency sub-band separately, with ISVC subspace dimension 100. The result is shown in Fig. 4. Fig. 5 shows normalized AID (after ISVC) subtracted from normalized SID. A comparison of Figs. 3 and 5 suggests that ISVC gives the biggest gain in region C. In fact, comparing Figs. 2 and 4, the average improvement of AID performance in this region is 24%, compared with average improvements of 6% and 2% in regions A and B, respectively, and an 12% decrease in region D. This suggests that ISVC is able to compensate for some of the speaker-dependent information in region C, which is noise from the perspective of AID, but not in region D.

V. Conclusion

This letter studied the effect of different frequency bands on automatic accent identification (AID) and speaker identification (SID) using the Accents of the British Isles speech corpus. Both the AID and SID systems were based on GMM-UBM approach. The competitiveness of our systems was demonstrated using full bandwidth (0–11.025 kHz) and band-pass filtered (0.23–3.4 kHz) speech experiments. The SID system achieved 100% accuracy using 30 second full bandwidth test signals and 512 component GMMs. The best AID accuracy was 60.34%, obtained using the band-pass filtered data, pitch-based SAD and 4096 component GMMs. The experimental results contrasting the utility of information in narrow sub-bands for the AID and SID tasks showed that it is useful to divide the spectrum into four regions: A (0 to 0.77 kHz), B (0.34 to 3.44 kHz), C (2.23 to 5.25 kHz) and D (3.40 to 11.02 kHz). Our experiments confirmed that speaker information dominates in regions A, corresponding to primary vocal tract resonance information, and D, corresponding to high-frequency sounds. In contrast, region B is most useful for AID, indicating that the vocal tract resonance information in this region is biased towards linguistic, rather than speaker information. Region C contains both types of information, although speaker information appears to dominate. When intersession variability compensation is applied to the AID system the biggest gain is observed in region C, where AID performance is improved by 24%, indicating that ISVC is able to factor out some of the speaker information in this region.

References

[1] L. Besacier, J. Bonastre, and C. Fredouille, “Localization and selection of speaker-specific information with statistical modeling,” *Speech Commun.*, vol. 31, pp. 89–106, 2000.

[2] R. C. Rose and D. A. Reynolds, “Text independent speaker identification using automatic acoustic segmentation,” in *IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 1990, vol. 1, pp. 293–296.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[4] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” *Speech Commun.*, vol. 17, pp. 91–108, 1995.

[5] A. Hanani, M. Russell, and M. Carey, “Human and computer recognition of regional accents and ethnic groups from British English speech,” *Comput. Speech Lang.*, vol. 27, no. 1, pp. 59–74, Jan. 2013.

[6] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation,” in *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2005, pp. 637–640.

[7] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, “A tutorial on text-independent speaker verification,” *EURASIP J. Appl. Signal Process.*, vol. 2004, pp. 430–451, 2004.

[8] D. R. Miller and J. Trischitta, “Statistical dialect classification based on mean phonetic features,” in *Int. Conf. Spoken Language Process. (ICSLP)*, 1996, vol. 4, pp. 2025–2027.

[9] L. M. Arslan and J. H. L. Hansen, “Language accent classification in American English,” *Speech Commun.*, vol. 18, pp. 353–367, 1996.

[10] C. Teixeira, I. Trancoso, and A. Serralheiro, “Recognition of non-native accents,” in *Eur. Conf. Speech Commun. and Technol. (EUSIPCO)*, 1997.

[11] M. Lincoln, S. Cox, and S. Ringland, “A comparison of two unsupervised approaches to accent identification,” in *Int. Conf. Spoken Language Process. (ICSLP)*, Sydney, Australia, 1998.

[12] R. Huang, J. H. L. Hansen, and P. Angkititrakul, “Dialect/accent classification using unrestricted audio,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 453–464, 2007.

[13] P. Angkititrakul and J. H. L. Hansen, “Advances in phone-based modeling for automatic accent classification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 634–646, 2006.

[14] M. A. Zissman, “Comparison of four approaches to automatic language identification of telephone speech,” *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 31–44, 1996.

[15] S. M. D’Arcy, M. J. Russell, S. R. Browning, and M. J. Tomlinson*, The Accents of the British Isles (ABI) Corpus*, pp. 115–119, 2004.

[16] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” in *A Speaker Odyssey: Speaker Recognition Wkshp.*, Crete, Greece, 2001, pp. 213–218.

[17] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, “Channel factors compensation in model and feature domain for speaker recognition,” in *IEEE The Odyssey: Speaker and Language Recognition Wkshp*, 2006, pp. 1–6.

[18] J. C. Wells*, Accents of English: Volume 2 – British Isles*. Cambridge, U.K.: Cambridge Univ. Press, 1982.