A Multi-Spectral Data Fusion Approach to Speaker Recognition

J. E. Higgins, R. I. Damper and C. J. Harris
Image Speech and Intelligent Systems Research Group
Department of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK.

Abstract *This paper describes a multi-spectral, multi-source approach to the important problem of speaker identification. The wideband speech signal is filtered into several sub-bands and the output time trajectory of each is individually modeled by linear prediction cepstral coefficients. These individual models are then matched against reference data and the scores combined using the sum rule of information fusion, before using a k-nearest-neighbor rule to decide the identified speaker. Multi-spectral processing is shown to deliver performance improvements over wideband recognition. The optimal number of filters is found to be 16. These results are interpreted in light of the hypothesis that the multi-spectral approach solves the bias/variance dilemma of commonly manifest in systems that are trained on example data.*

Keywords: speaker recognition, feature-set construction, multi-spectral fusion

1 Introduction

Automatic speaker recognition (ASR), whereby a computer attempts to recognize an individual from their voice, is an important, emerging technology with many potential applications in commerce and business, security, surveillance etc. This paper is concerned with the application of modern data engineering techniques to the problem of ASR. The main idea here is the use of a multi-spectral approach, in which the wideband acoustic signal is pre-processed by a bank of bandpass filters to give a set of time-varying outputs – so-called *sub-band* signals. Because these time trajectories vary slowly relative to the wideband signal, the problem of representing them by some data model should be simplified. A major goal for this paper is to test if this is so, and if so, to determine the optimal number of sub-bands. Since we now have several time trajectories to consider rather than just one, the question arises of how to (re)combine or fuse the information in each, to reach an overall decision about speaker identity.

The remainder of the paper is organized as follows. Section 2 provides a motivation for research into recognition. Section 3 introduces the multi-spectral aspect of the recognition system and includes fuller discussion on the possible benefits to an identification system. In section 4, the component parts of the baseline multi-spectral system which provides the foundation for this research are described in turn. Finally, section 6 concludes with discussion of the issues raised by multi-spectral recognition and some possible avenues of future work.

2 Speaker Recognition

Recognition can be divided into speaker *verification* and speaker *identification* tasks, each of which may in turn be *text-independent* or *text-dependent* [1, 2]. In verification, there is an *ab initio* claim about speaker identity, and the aim is to determine if a given utterance was produced by the claimed speaker. This is done by testing the model of the claimed speaker against the utterance, comparing the score to a threshold, and deciding on the basis of this comparison whether or not to accept the claimant. In identification, there is no *ab initio* identity claim, and the system must typically decide who the person is, or that the person is unknown.

In text-independent recognition, there are no limits on the vocabulary employed by speakers. This is in contrast to text-dependent recognition, where the presented utterance must be from a set of predetermined words or phrases. As text-dependent recognition only models the speaker for a limited set of phonemes in a fixed context, it generally achieves higher recognition rates than text-independent recognition, which must model a speaker for a variety of phonemes and contexts.

Speaker recognition is an example of biometric personal identification [3]. Biometric techniques based on intrinsic characteristics (such as voice, finger prints, retinal patterns) have an advantage over artifacts for identification (keys, cards, passwords) because biometric attributes cannot be lost or forgotten. Biometric techniques are generally believed to offer a reliable method of identification, since all people are physically different to some degree. Automatic speaker identification and verification are often considered to be the most natural and economical methods for avoiding unauthorized access to physical locations or computer systems [1]. Thanks to the low cost of microphones and the universal telephone network, the only cost for a speaker recognition system may be the software.

In this paper, we are primarily interested in text-dependent identification. Success depends on extracting and modeling the speaker-dependent characteristics of the speech signal which can effectively distinguish one talker from another.

Figure 1 shows the structure of a typical, simple identification system. In general, identification consists of five steps:

- digital speech data acquisition
- feature extraction
- pattern matching
- identification decision
- enrollment to generate reference models of each speaker

Initially, the acoustic sound pressure wave from an unknown speaker is transformed into an analog signal by a microphone or telephone handset. The analog signal is then passed through an anti-aliasing filter before being sampled to form a digital signal by an analog-to-digital converter.

In feature extraction, each frame of speech (typically spanning 10–30 ms of the speech waveform) is mapped into a multidimensional feature space creating a sequence of feature vectors $\mathbf{x}_i$. This sequence is compared to existing speaker models, created during the enrollment step, by pattern matching, resulting in a match score z_i for each of the speaker models. The match score gives an indication of the similarity between the sequence of vectors and the models of all the known speakers. The last step consists of a decision as to speaker identity. Before use, speakers must *enroll* on the system. During enrollment, speaker models are created for all authorized users and stored for later reference during identification.

3 Multi-Spectral Processing

In a seminal and influential paper, Allen [4] popularized the earlier notion of Harvey Fletcher that the decoding of speech signals by humans is based on decisions in narrow frequency bands that are processed independently of each other. Decisions from these frequency bands are combined such that the global error rate is equal to the product of the band-limited error rates within the independent frequency channels. This means that if any frequency band yields a zero (or low) error rate then the resulting global error rate would also be zero (or very low), regardless of the error rates of the remaining bands. While this has come to be known as the Fletcher-Allen principle, Allen himself refers to it as “the Stewart-Fletcher multiindependent channel model” (p. 572). He further characterizes the approach as “across-time” rather than the more usual “across-frequency” processing (p. 575) typified by template matching in automatic speech recognition. In this paper, we will refer to this as *multi-spectral processing.*

The positive benefits of this new approach to speech recognition are starting to be investigated

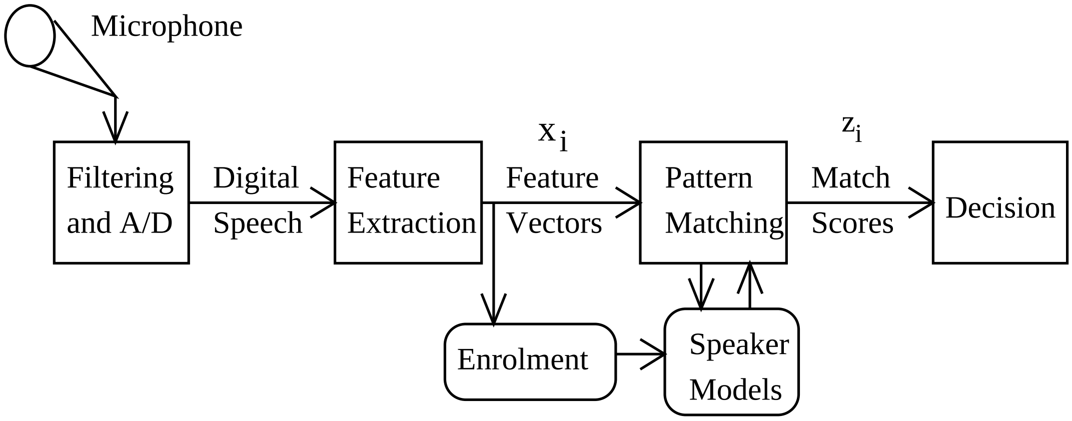

Figure 1: Block diagram of a typical speaker-identification system.

and reported [5, 6, 7, 8]. There are several cogent reasons why it might also profitably be applied to *speaker* recognition:

- The deleterious effect of narrow-band noise may be reduced. If noise only affects some frequency bands, then the remaining (clean) bands should carry sufficient information to allow the correct decision still to be reached. This follows from the (idealized) Fletcher-Allen principle according to which only one error-free band is required for correct recognition.

- Certain bands may contain more speaker-specific information than others. Weighting these to emphasize their contribution to the overall score should lead to better recognition rates. In fact, some bands might be better for some speakers than others, so that speaker-specific weighting during the information fusion – or (re)combination – stage may be possible. Note, however, that this assumes a form of fusion in which weighting can be sensibly done. (If, for instance, combination is by multiplication of scores, then weighting has no effect.)

- Successful recognition critically depends on building good speaker models from the training data. Data modeling, however, is subject to the well-known *bias/variance dilemma* [9]. According to this, models with too many adjustable parameters (relative to the amount of training data) will tend to overfit the data, exhibiting high variance, and so will generalize poorly. On the other hand, models with too few parameters will be over-regularized, or biased, and so will be incapable of fitting the inherent variability of the data. Multi-spectral processing offers a practical solution to the bias/variance dilemma by replacing a large, unconstrained data modeling problem by several smaller (and hence more constrained) problems. Empirical support for this notion in the specific context of speaker recognition comes from the work of Reynolds [10], who writes: “giving too much spectral resolution will degrade performance by modeling spurious spectral events or introducing too many parameters to be trained” (p. 642).

There are, however, several practical issues to be resolved before these advantages might be realized:

- The number, width and location of the frequency bands must be optimized. Sub-bands designed for speech recognition may not be suitable for speaker recognition: it may be that the frequency division should best be done on a speaker-specific basis for speaker recognition.

- Some knowledge is required of which bands contain the most speaker-dependent information. The scores from these bands might then be emphasized to improve recognition.

- The features to be used for recognition must be decided. Again, features designed for speech recognition may not be suitable for speaker recognition [2]. It is also possible that features which are appropriate for wide-band speaker recognition are less so for multi-spectral processing.

To date, relatively few workers have studied this problem. In the conference literature, [11], [12], [13] and [14] have all presented empirical results which confirm that worthwhile performance advantages can be gained from multi-spectral processing in speaker recognition. Taken together, however, these prior works do not cover anything like the full range of implementation options, so that many of the aforementioned questions remain open. Further, there is still only a rudimentary understanding of multi-spectral processing – and precisely how it delivers performance improvements – from a theoretical perspective.

4 Identification System

This section describes the different components that make up the identification system.

4.1 Database

The text-dependent Millar database from British Telecom was specifically designed and recorded for text-dependent speaker recognition studies. It consists of 43 male and 14 female native English speakers saying the digits *one* to *nine*, *zero*, *nought* and *oh* 25 times each. Recordings were made in five sessions spaced over three months, to capture the variation in speakers‘ voices over time which is one of the most important aspects of speaker recognition [15]. The speech was recorded digitally in a quiet environment using a high-quality microphone, and a sampling rate of 20 kHz with 16 bit resolution. The database was also made available at an 8 kHz sampling rate. In this version, the speech has been band-passed to telephone quality and then downsampled. Only this latter version was used.

For the experiments, 12 male speakers were used saying the word *seven*. The first two sessions (i.e. 10 repetitions of *seven*) were used as references and the remaining three sessions (15 repetitions) were used for testing.

4.2 Sub-Band Processing

The wideband signal was split into various numbers of sub-bands. Filters were simple second-order Butterworth, spaced on the psychophysical mel scale [16], covering the frequency range up to 3,600 Hz. There are many possible features that can be extracted from a speech signal, e.g. fundamental frequency, formant frequencies, and linear predictor (LP) coefficients. For recognition purposes, it is important to use a feature set that maximally discriminates between speakers. In this research, the feature set is based on cepstral coefficients. Cepstral analysis is motivated by, and was designed for, problems centered on *voiced* speech [17] but also works well for unvoiced sounds. Cepstral coefficients have been used extensively as the features in speaker recognition [18, 19]. This is because a simple recursive relation (see below) can be used to transform the LP coefficients into cepstral coefficients.

The time trajectories in each sub-band were modeled using an analysis frame of 20 ms, Hamming windowed and overlapping by 50%, and 12th order linear prediction [20]. These were then used to create cepstral coefficients via the recursion described by Atal [21]. That is, the LP cepstrum (or pseudo-cepstrum) is used, rather than the original (power or complex) cepstrum which would be obtained from Fourier analysis.

4.3 Pattern Matching

A popular method of pattern matching in speaker recognition systems uses ‘templates’. The input signal is represented as a series of feature vectors that characterize the speech of a particular

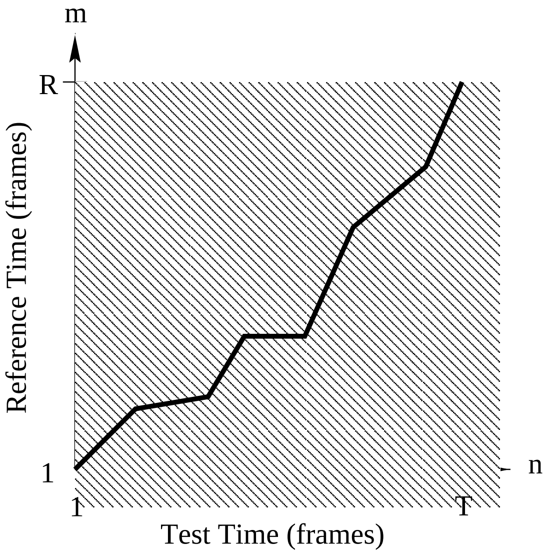

Figure 2: Typical DTW plot, illustrating the optimal warp path mapping the test time axis n into the reference time axis m.

speaker [22]. This time-ordered set of features constitutes the template. Even if the same speaker utters the same word on different occasions, the duration changes each time with nonlinear expansion and contraction. Therefore, any template matching algorithm needs to be able to cope with this: we use the popular technique of dynamic time warping (DTW) because of its ability to handle nonlinear time scale variations. It combines alignment and distance computation through a dynamic programming procedure [23]. It is normal to use the Euclidean distance measure when working with cepstral coefficients. Figure 2 depicts the DTW procedure schematically.

4.4 Fusion

Kittler, Hatef, Duin, and Matas [24] recently developed a common theoretical framework for combining classifiers which use distinct pattern representations. They outlined a number of possible combination schemes such a product, sum, min, max, and majority vote rules, and compared their performance empirically using two different pattern recognition problems. Kittler et al. found that the sum rule outperformed the other classifier combination schemes. This surprised them, because the statistical assumptions underlying this rule are stronger than, say, those for the product rule and it is clear that these assumptions do not hold well.

To explain this empirical finding, they investigated the sensitivity of various schemes to estimation errors. Their analysis showed that the sum rule is the most resilient to estimation errors, so almost certainly explaining its superior performance. Accordingly, the sum rule is used, at least initially, for combination purposes in this research while recognizing that this is one area which could benefit from further research by investigating other rules and methods of combination.

4.5 Decision Rule

There are 15 test utterances per speaker, each of which is matched to the 10 reference utterances for all 12 speakers – a total of 120 comparisons. These are then ranked (closest matches first) and the k-nearest-neighbor rule applied with $k = 5$. That is, the speaker maximally represented among the top five ranking matches is declared to be the identified person.

5 Results

To investigate the benefits of multi-spectral processing, as well as answering the question of the optimal number of sub-bands, we have collected identification results as the number of filters varies from 2 to 24. For comparison, recognition was performed using the wideband (unfiltered) speech signal also. Figure 3 displays the results.

It is clear that a multi-spectral recognition system can perform better than one using just the wideband signal. Using the wideband spectrum, the system achieved 85% recognition rate. By contrast, the best-performing multi-spectral system, using 16 mel-spaced sub-bands, produced a recognition rate of 96%. This is a very considerable improvement.

Using a small number of filters (< 6)), performance was generally worse than the wideband system. The reason for this is currently unknown, but we conjecture that too much spectral energy is removed by the filterbank, i.e. the regions of overlap between adjacent filters are too wide. Conversely,

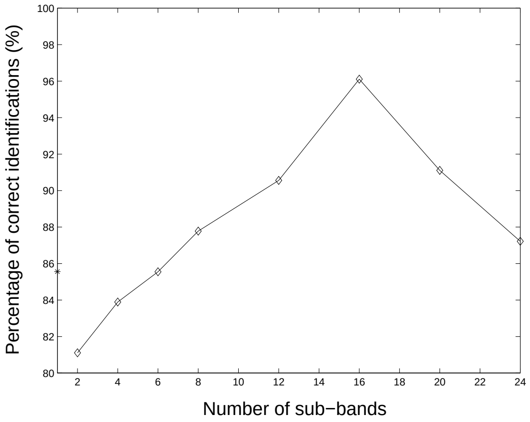

Figure 3: Percentage of correct identifications for different numbers of mel-spaced sub-bands (* indicates wideband).

it is possible to have too many filters. Performance reduces when there at 20 filters or more. We attribute this to attempting to fit too many parameters in the data models describing each speaker.

From the perspective of time-frequency duality, it seems intuitively reasonable that there should be some such trade-off. With a small number of filters, we will be attempting to fit the time trajectories too closely, having only a few parameters to do so. With a large number of filters, we will be attempting to fit the frequency distribution too closely but with more parameters than can be reliably estimated from the data. There is an interesting convergence with Allen's comment [4]: "It has been reported ... that 10 bands is too few, and 30 bands gives no improvement in accuracy over 20" (p. 572).

6 Discussion and Conclusions

The results highlight the advantage of a using multi-spectral approach to speaker recognition. We believe that the approach offers a practical solution to the bias/variance dilemma manifest in trainable systems, and so leads to improved data modeling. The problem of fitting parameters to training data is constrained by requiring them to be more or less uniformly deployed across frequency. Although multi-spectral processing increases performance, there is a limit to how many sub-bands can be used before performance starts to decrease. Here, it seems that 16 is the optimal number. This finding is interpreted in data-modeling terms as reflecting an attempt to fit too many parameters for the available training data. By contrast, the wideband approach (or use of a small number of filters) attempts data modeling with too few, unconstrained parameters.

The traditional approach to identification has been to base the development of recognition systems on *a priori* knowledge. The prior knowledge has been applied to such things as choosing the type and number of feature parameters and determining the pattern matching method to use. Current speaker identification systems produce reasonable results but still lack the necessary performance if they are to be used routinely by the general public. Furui has listed 16 open questions about speaker recognition which need to be addressed if performance is to be improved. One of these concerns the selection of feature parameters: commonly cepstral (or delta cepstral) coefficients. These are employed principally (or only) because they are familiar from their use in speech recognition. Hence, they may not optimally discriminate between different speakers. From this perspective, there seems much to be gained from automatic (data-driven) selection of features – and other architectural parameters.

Future work will look at possible ways of implementing a data-driven strategy for number and placement of the filters, and for automatically determining the type and number of features to be used in each sub-band. We will also explore other combination schemes and will extend the work to speaker verification. Finally, we propose a direct test of our hypothesis of improved data modeling, by varying the number of parameters fitted in the different filtering scenarios.

7 Acknowledgements

The authors would like to thank Robert Finan for kindly providing some of his code and the UK Engineering and Physical Sciences Research Council for their support.

References

[1] J.P. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1436–1462, Sep 1997.

[2] S. Furui. Recent advances in speaker recognition. *Pattern Recognition Letters*, 18(9):859–872, September 1997.

[3] G.R. Doddington. Speaker recognition – identifying people by their voices. *Proceedings of the IEEE*, 73(11):1651–1664, Nov 1985.

[4] J. B. Allen. How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577, October 1994.

[5] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *Proceedings of International Conference on Spoken Language Processing (ICSLP '96)*, pages 426–429, Philadelphia, PA, 1996.

[6] H. Hermansky, S. Tibrewala, and M. Pavel. Towards ASR on partially corrupted speech. In *Proceedings of 4th International Conference on Spoken Language Processing, ICSLP 96*, volume 1, pages 462–465, Philadelphia, PA, 1996.

[7] S. Tibrewala and H. Hermansky. Subband based recognition of noisy speech. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 1255–1258, 1997.

[8] H. Hermansky and S. Sharma. TRAPS – Classifiers of temporal patterns. In *Proceedings of 5th International Conference on Spoken Language Processing, ICSLP 98*, Sydney, Australia, 1998. Paper 615 on CD-ROM.

[9] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

[10] D. A. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):639–643, October 1994.

[11] L. Besacier and J. Bonastre. Subband approach for automatic speaker recognition: Optimal division of the frequency domain. In *Proceedings of 1st International Conference on Audio- and Visual-Based Person Authentication (AVBPA)*, Crans-Montana, Switzerland, 1997.

[12] R. Auckenthaler and J. S. Mason. Equalizing sub-band error rates in speaker recognition. In *European Speech Communication Association (ESCA) Conference, Eurospeech '97*, pages 2303–2306, Rhodes, Greece, 1997.

[13] P. Sivakumaran, A. M. Ariyaeeinia, J. A. Hewitt, and J. A. Malcolm. An effective sub-band based approach for robust speaker verification. *Proceedings of the Institute of Acoustics*, 20(6):69–72, 1998.

[14] P. Sivakumaran, A. M. Ariyaeeinia, and J. A. Hewitt. Sub-band speaker verification using dynamic recombination weights. In *Proceedings of 5th International Conference on Spoken Language Processing, ICSLP 98*, Sydney, Australia, 1998. Paper 1055 on CD-ROM.

[15] R.A. Finan, A.T. Sapeluk, and R.I. Damper. Imposter cohort selection for score normalisation in speaker verification. *Pattern Recognition Letters*, 18(9):881–888, September 1997.

[16] E. Zwicker and E. Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, 68(5):1523–1525, 1980.

[17] J.R. Deller, J.G. Proakis, and J.H.L. Hansen. *Discrete-Time Processing of Speech Signals.* Macmillan, New York, 1993.

[18] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transac-*

tions on Acoustics, Speech and Signal Processing, ASSP–29(2):254–272, April 1981.

[19] D.A. Reynolds and R.C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, Jan 1995.

[20] J. D. Markel and A. H. Gray. *Linear Prediction of Speech*. Springer-Verlag, Berlin, Germany, 1976.

[21] B.S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustic Society of America*, 55(6):1304–1312, June 1974.

[22] J.M. Naik. Speaker verification: A tutorial. *IEEE Communications Magazine*, pages 42–48, January 1990.

[23] D. O'Shaughnessy. *Speech Communication: Human and Machine*. Addison-Wesley, 1987.

[24] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.