Schedule for ELEC9781
Forensic Voice Comparison and the Evaluation of Evidence
–
Second part

Thursday 25 August 2011 (Week 6)

Voice activity and feature extraction for speaker characterization – acoustic

Campbell (1997) (Sections I, IIA-C); Bimbot et al. (2004) (section 1); Kinnunen and Li (2010) (section 1);
Optional (history): Rosenburg (1976); Doddington (1985)

- Explain the components of a speaker recognition system

- What is the objective of speaker verification ? How are speaker recognition systems evaluated ? How is this different to forensic voice comparison ?

- Over what duration of time are features normally extracted in speaker recognition ? How does this compare with forensic voice comparison ?

Atal and Rabiner (1976); Kinnunen and Li (2010) (section 5.1); Epps and Ambikairajah (2011) (section 3.2.1); Benyassine et al. (1997).
Optional (further reading): Savoji (1989); Shen et al. (1998);

- What are the main components of voice activity detectors ?

- Name as many methods as you can for automatically determining whether a signal segment comprises speech or non-speech

- What are the main challenges for accurate voice activity detection ? When is VAD most likely to be inaccurate ?

- Comment on how VAD is used in speaker recognition systems. For example, do researchers quote the accuracy of their VAD ? How do they choose their VAD thresholds ?

Rabiner et al., (1976); Try to absorb what you can from the following: Talkin (1995); Welling and Ney (1998); Snell and Milinazzo (1993).

- How can the fundamental frequency be automatically extracted from a voiced speech signal ? (investigate one simple, widely used method)

- Explain the types of errors that occur and why

- How can the formant frequencies be automatically extracted from a voiced speech signal ? (investigate one simple, widely used method)

- Explain the types of errors that occur and why

- Suppose someone recommends a pitch or formant estimation tool to you. Explain how you would validate it before applying it to your system.

- Why are pitch and formants not used much in state of the art speaker recognition systems ?

Thursday 1 September 2011 (Week 7)

Feature extraction for speaker characterization – acoustic, prosodic, linguistic

Davis and Mermelstein (1980); Bimbot et al. (2004) (section 2.1); Epps and Ambikairajah (2011) (section 3.2.2.2)

- Explain the purpose of each of the stages of MFCC extraction

Sambur (1975); Epps and Ambikairajah (2011) (section 3.2.2); Kinnunen and Li (2010) (sections 2.1, 3.1-3.3);
Optional (further reading): Atal (1974); Davis and Mermelstein (1980); Reynolds (1994); Bimbot et al. (2004) (section 2.2)

- What are the desirable attributes of a feature for speaker recognition ?

- Select two acoustic features and prepare to present them to the class. Try to explain (i) what information it captures, (ii) how it is extracted, and (iii) how it performs relative to other features (and why, if you can)

Reynolds et al. (2003); Kockmann et al., (2010); Epps and Ambikairajah (2011) (sections 3.2.4-3.2.5); Kinnunen and Li (2010) (sections 3.4-3.5)

- Select one prosodic or linguistic feature and prepare to present it to the class (continued from previous week). Try to explain (i) what information it captures, (ii) how it is extracted (in as much detail as you can), and (iii) how it performs relative to other features (and why, if you can)

- Comment on differences between acoustic, prosodic and linguistic features, and their relative performances

Thursday 8 September 2011 (Break week)

Feature extraction for speaker characterization – speaker discrimination and dynamic variants

Lu and Dang (2008); Thiruvaran et al. (2009);
Optional (related reading): Campbell (1997) (p 1447-1448)

- What is the motivation of the authors in the first two papers ?

- Comment on their methodology – which methods would you choose and why ?

- Comment on what they find – does it match their hypotheses, or is it surprising ?

- Try to form your own opinion of the relative importance of different frequency bands for speaker discrimination and be ready to justify it to the class

Reynolds (1994); Epps and Ambikairajah (2011) (section 3.2.3)

- Try to name as many methods as you can for capturing dynamic information in acoustic features

- Explain how RASTA works and its purpose

Sönmez et al. (1998); Kockmann and Burget (2008); Kockmann et al. (2010) (in particular Sections 2.2-2.3).
Optional (reference reading for HMMs): Rabiner (1989)

- Explain some different approaches to parameterising feature contours. Explain the process by which you would choose a parameterization, including the number of parameters.

- Explain at least two ways in which hidden Markov models could be used to parameterise a sequence of feature values

Epps and Ambikairajah (2011) (section 3.2.3)

- Discuss the trade-offs to consider when deciding how to represent the dynamic information contained in a feature

Thursday 15 September 2011 (Week 8)

Normalisation, speaker modeling and Gaussian mixture models

Mammone et al. (1996) (from "Cepstral derivatives" to the end of "Cepstral mean subtraction"); Pelecanos and Sridharan (2001); Garcia et al. (2006) (just try to understand their basic idea)

- Explain the purpose of feature normalization

- Explain in detail methods for achieving feature normalization. Compare these and try to explain differences in their performance.

Reynolds and Rose (1995); Reynolds et al. (2000); Bimbot et al. (2004) (section 3.2); Kinnunen and Li (2010) (section 4)

- Explain the process by which k-means clustering (or vector quantization codebook design) is performed

- Compare and contrast clusters obtained by k-means with individual Gaussian mixtures

- How are k-means and Gaussian mixture models related to vector quantization ?

- What are the differences between vector quantization and Gaussian mixture models ?

- Research the process by which GMM parameters are usually estimated. Find the update equations for GMM parameters.

- Compare and contrast *in detail* the k-means algorithm with the process by which GMM parameters are estimated.

- How many GMM mixtures are typically used ? How is this determined ?

Thursday 22 September 2011 (Week 9)

Speaker modeling, Gaussian mixture models and adaptation

Bimbot et al. (2004) (section 3.3); Reynolds and Rose (1995); Reynolds et al. (2000); Kinnunen and Li (2010) (sections 4.1-4.2)

- What is meant by the term 'likelihood' ?

- How should a universal background model be generated ? Can you find guidelines for how to select UBM data ? (what about for forensic purposes?)

- Explain the process of MAP adaptation

- Name two advantages of MAP-adapted models over GMM models without adaptation

- Explain the purpose of the relevance factor

- Are there other methods for determining an adapted GMM model set ?

Thursday 29 September 2011 (Week 10)

Lab: Putting it all together – features and modeling – See Laboratory Exercise Sheet

- Familiarize yourself with VoiceBox

- Familiarize yourself with HTK

- Create synthetic data in MATLAB, then write features to a file in HTK format. Train a GMM and use it to calculate a likelihood.

- Implement VAD, feature extraction, then extract features from speech data and write the features to a file in HTK format

- Train a GMM and use it to calculate a likelihood

- If time permits, try MAP adaptation

- Experiment with different parameter choices

Thursday 6 October 2011 (Week 11)

Speaker modeling (continued)

Kua et al. (2011)

- MFCCs have long been recognized as the most effective feature for speech recognition (where speaker identity is a source of variability). How is it that MFCCs could also be the feature of choice for speaker recognition systems ?

- In acoustic-phonetic FVC, speech data are segmented to identify specific tokens (e.g. phonemes, words) for comparison (i.e. to compare 'like with like'). Comment on the claim that speaker recognition systems use speech indiscriminately in this respect.

Supervectors

Gauvain and Lee (1994) (section 3); Fauve et al. (2007) (section IV); Reynolds (1997); Bimbot et al. (2004); Campbell et al. (2005); Kinnunen and Li (2010) (sections 6.1-6.5, 6.8)

- What was the motivation for introducing supervectors ? If you can't find it in the papers, try to guess. Are there any other advantages of supervectors ?

- List the different types of supervectors.

- Explain how each captures speaker-specific information.

- What is the advantage of working with supervectors as opposed to e.g. features, or GMM parameters ?

Thursday 13 October 2011 (Week 12)

Channel compensation

Campbell et al. (2006); Fauve et al. (2007) (sections V, VI, VII); Kenny and Dumouchel (2004); Kenny et al. (2008); Kinnunen and Li (2010) (sections 6.6-6.8)

- Be ready to explain to the class how principal component analysis works.

- Sketch the effect of nuisance attribute projection (NAP) on a small set of synthetic two-dimensional data.

- Write down MATLAB commands to implement NAP.

- Explain the concept of joint factor analysis.

- Sketch the procedure by which JFA training and testing is conducted.

- Discuss the suitability of channel compensation methods for the forensic voice comparison application.

Thursday 20 October 2011 (Week 13)

Segment selection for speaker characterization

Vair et al. (2007); Bocklet and Shriberg (2009); Pruthi and Espy-Wilson (2004)

- What is the generic hypothesis underlying the use of segment selection methods ?

- What are the main components of segment selection methods ?

- List as many different kinds of automatic segment selection methods as you can from the literature. If you can, comment on the differences between manual and automatic segmentation methods.

- List some types of segments you think may be helpful for forensic voice comparison, and explain your choices.

Rabiner (1989)

- Explain what types of signals an HMM can model.

- What is the difference between a GMM and an HMM ?

- What are the parameters of an HMM ? Briefly explain how these are estimated.

- Why are HMMs used for segmentation ? What is meant by the term 'forced alignment' ? (you may need to consult other references for this)

- Explain in detail how you would construct a system to recognize different phonemes.

Acknowledgements

Thanks are due to Jia Min Karen Kua for helpful discussions on channel compensation.

Bibliography

Where papers have not been made available via http://forensic.unsw.edu.au/ELEC9781.html, they can be downloaded from http://ieeexplore.ieee.org/Xplore/guesthome.jsp?reload=true (must be on UNSW network to access) or from http://scholar.google.com.au/.

Atal, B. S., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *JASA*, vol. 55, 1974, pp. 1304-1312.

Atal, B. S., and Rabiner, L. R., "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition", *IEEE Trans. Acoust., Sp. and Sig. Proc.*, vol. ASSP-24, no. 23, 1976, pp. 201-212.

Benyassine, A., Shlomot, E., Su, H.-Y., Massaloux, D., Lamblin, C., and Petit, J.-P., "ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications", *IEEE Communications Magazine*, vol. 35, no. 9, 1997, pp. 64-73.

Bocklet, T. and Shriberg, E., "Speaker recognition using syllable-based constraints for cepstral frame selection", in *Proc. IEEE ICASSP*, 2009, pp. 4525-4528.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska, D., and Reynolds, D. A., "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing, Special issue on biometric signal processing*, 2004.

Campbell, J. P., "Speaker recognition: A tutorial", *Proceedings of the IEEE*, vol. 85, no. 9, Sept. 1997, pp. 1437-1462.

Campbell, W. M., Sturim, D. E., and Reynolds, D. A., "Support vector machines using GMM supervectors for speaker verification", *IEEE Signal Processing Letters*, vol. 13, no. 5, 2005, pp. 308-311.

Campbell, W., Sturim, D. E., Reynolds, D., and Solomonoff, A., "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. IEEE ICASSP*, 2006, pp. 97-100.

Davis, S. B., and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-28, 1980, pp. 357-366.

Doddington, G. R., "Speaker recognition – Identifying people by their voices", *Proceedings of the IEEE*, vol. 73, no. 11, 1985, pp. 1651-1664.

Epps, J., and Ambikairajah, E., "Speech characterization and feature extraction for speaker recognition", in Li, G, and Li, H. (eds), *Advanced Topics in Biometrics*, World Scientific, 2011, to appear.

Fauve, B. G. B., Matrouf, D., Scheffer, N., Bonastre, J. F., and Mason, J. S. D., "State-of-the-art performance in text-independent speaker verification through open-source software", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1960–1968, 2007.

Ferrer, L., Scheffer, N., and Shriberg, E., "A comparison of approaches for modeling prosodic features in speaker recognition", in *Proc. IEEE ICASSP*, pp. 4414–4417, Mar. 2010.

Garcia, L., Segura, J. C., Ramirez, J., de la Torre, A., and Benitez, C., "Parametric Nonlinear Feature Equalization for Robust Speech Recognition," in *Proc. IEEE ICASSP*, vol. I, 2006, pp. 529-532.

Gauvain, J., and Lee, C., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639-643, 1994.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P., "RASTA-PLP speech analysis technique", in *Proc. IEEE ICASSP*, vol. 1, 1992, pp. 121-124.

Kenny, P., and Dumouchel, P., "Experiments in speaker verification using factor analysis likelihood ratios", in *Proc. Odyssey*, 2004, pp. 219-226.

Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P., "A study of inter-speaker variability in speaker verification", *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, July 2008. Available: http://www.crim.ca/perso/patrick.kenny

Kinnunen, T., and Li, H., "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, 2010, pp. 12-40.

Kockmann, M. and Burget, L. "Contour Modeling of Prosodic and Acoustic Features for Speaker Recognition", in *Proc. of Spoken Language Technology*, 2008, pp. 45-48.

Kockmann, M., Burget, L., and Cernocky, J., "Investigations into prosodic syllable contour features for speaker recognition",in *Proc. IEEE ICASSP*, Mar. 2010, pp. 4418–4421.

Kua, J. M. K., Epps, J., Nosratighods, M., Ambikairajah, E., and Choi, E. H. C., ″Using Clustering Comparison Measures for Speaker Recognition″, to appear in *Proc. IEEE Int. Conf. on Acoust., Speech and Sig. Proc.* (Prague, Czech Republic), 2011.

Lu, X., and Dang, J., "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech Communication*, vol. 50, 2008, pp. 312 – 22.

Mammone, R., Zhang, X., and Ramachandran, R. P., "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, 1996, pp. 58-71.

Pelecanos, J., and Sridharan, S., "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey* (Crete, Greece), 2001, pp. I – 640.

Pruthi, T., and Espy-Wilson, C. Y., "Acoustic parameters for automatic detection of nasal manner", *Speech Communication*, vol. 43, no. 3, August 2004, pp. 225-239.

Rabiner, L. R., Cheng, M. J., Rosenberg, A. E., and McGonegal, C. A., "A comparative performance study of several pitch detection algorithms", *IEEE Trans. on ASSP*, vol. ASSP-24, no. 5, 1976, pp. 399–418.

Rabiner, L. R., "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE*, vol. 77, no. 2, 1989, pp. 257-286

Reynolds, D. A., "Experimental evaluation of features for robust speaker identification", *IEEE Trans. Speech and Audio Proc.*, vol. 2, 1994, 639-643.

Reynolds, D. A. and Rose, R. C., "Robust text-independent speaker identification using Gaussian mixture models", *IEEE Trans. Acoust., Speech and Signal Processing,* vol. 3, no. 1, 1995, 72–83.

Reynolds, D. A., "Comparison of background normalization methods for text-independent speaker verification", in *Proc. European Conference on Speech Communication and Technology*, vol. 2, September 1997, pp. 963-966.

Reynolds, D. A., Quatieri, T. F., Dunn, R. B., "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing Special Issue on NIST 1999 Speaker Recognition Workshop*, vol. 10, no. 1-3, 2000, pp. 19-41.

Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., and Xiang, B., "The SuperSID project: exploiting high-level information for high accuracy speaker recognition", in *Proc. IEEE ICASSP*, 2003, pp. 4:784–787.

Rosenburg, A. E., "Automatic speaker verification: A review", *Proc. IEEE*, vol. 64, no. 4, April 1976, pp. 475-487.

Sambur, M. R., "Selection of acoustic features for speaker identification", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, 1975, pp. 176–182.

Savoji, M. H., "A robust algorithm for accurate endpointing of speech signals", *Speech Communication*, vol. 8, 1989, pp. 45-60.

Shen, J.-L., J.-W. Hung and L.-S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proc. Int. Conf. on Spoken Language Processing*, 1998.

Snell, R. C., and Milinazzo, F., "Formant location from LPC analysis data", *IEEE Trans. Speech Audio Processing*, vol. 1, Apr. 1993, pp. 129-134.

Sönmez, K., Shriberg, E., Heck, L., and Weintraub, M., “Modeling dynamic prosodic variation for speaker verification”, in *Proc. Int. Conf. on Spoken Lang. Process.*, 1998, pp. 3189-3192 .

Talkin, D., “A robust algorithm for pitch tracking (RAPT)”, in Kleijn, W. & Paliwal, K. (eds), *Speech coding and synthesis*, Elsevier, New York, 1995, pp. 495-518.

Thiruvaran, T., Ambikairajah, E., and Epps, J., "Analysis of band structures for speaker-specific information in FM feature extraction", In *Proc. INTERSPEECH*, 2009, pp. 1111-1114.

Welling, L. and Ney, H., “Formant Estimation for Speech Recognition”, *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, 1998, pp. 36–48.

Vair, C., Colibro, D., Castaldo, F., Dalmasso, E., and Laface, P., “Loquendo - Politecnico di Torino’s 2006 NIST speaker recognition evaluation system,” in *Proc. INTERSPEECH*, 2007, vol. 1, pp. 113 – 116.