An Accurate and Robust Gender Identification Algorithm

Andrea DeMarco, Stephen J. Cox

School of Computing Sciences, University of East Anglia, Norwich, England
a.de-marco@uea.ac.uk, s.j.cox@uea.ac.uk

Abstract

We describe a robust, unsupervised method of automatic gender identification from speech. We first design a baseline gender classifier based on MFCC features, and add a second classifier that uses context-dependent but text-independent pitch features. The results of these classifiers are then examined for disagreements in gender classification. Any disagreements are resolved by the use of a novel pitch-shifting mechanism applied to the utterances. We show how the acoustic context classifier provides very good gender identification results, and how these are further enhanced by the pitch-shifting process. Furthermore this enhancement is preserved across a set of different corpora.

Index Terms: gender identification, speaker recognition, pitch

1. Introduction

The problem of automatic gender identification in speech has been studied using various techniques. Wu and Childers [1] use various features (autocorrelation, linear prediction, cepstrum and reflection) extracted from clean speech. They claim that these different features can all effectively be used for gender identification. In another study by Pronobis and Magimai [2] the focus is solely on pitch and cepstral features. They claim that these work equally well for clean speech, but that cepstral features give better gender classification in adverse conditions. Also high order cepstral features and spectral dynamics give more robust results on mismatched training and test data.

Tran and Sharma [3] propose an automatic gender identification algorithm based on building separate Hidden Markov Models (HMMs) for the genders. This work makes the assumption that speakers in the training and testing sets have a closed vocabulary that they can use for utterances. With a closed vocabulary it is possible to construct a HMM for each gender based on the sequences of observations in the training set. In the test case, the utterance is then matched against both gender HMMs, and the HMM that gives the highest score is selected. Low error rates were reported in this experiment (2.4% for male speakers and 6.1% for female speakers). The main problem with this approach however is that in normal conversational speech, the vocabulary is virtually unlimited, thus making gender identification systems built on closed vocabulary HMMs impractical. On the other hand, this work shows that knowing the context of a sound (via HMM states, in this case) has a strong impact on the performance of a gender identification system. In this approach, training was performed on a relatively low number of test samples, from a low number of speakers (8 males and 8 females).

Zeng et. al [4] describe a novel Gaussian Mixture Model (GMM) classifier based on a concatenation of pitch values with the corresponding RASTA-PLP feature vector. A small order GMM (4-8 components) is sufficient in their experiments, and the performance is very robust in the presence of speech data from different corpora and languages. Other features and classification methods have been proposed for the gender classification problem, with results reaching 95% accuracy for the test systems [5, 6, 7].

Our approach is designed to find a close acoustic context to the content that is being analyzed in speech using predefined acoustic templates built by MFCC codebooks. Once the acoustic context is determined, the pitch information expected within that acoustic context is compared to male and female pitch templates and a gender decision is made. Furthermore, we exploit inconsistencies between gender classifiers by looking at the effect of pitch-based distortions of the original speech signal to give a refined classification where possible.

2. Gender classification methodology

2.1. Baseline classification

In our baseline classifier, we use MFCC feature vectors extracted from continuous speech, either from the TIMIT [8] or from the ABI [9] corpora. Two different vector quantizer models are built, one per gender, by clustering (K-Harmonic Means) the MFCC feature vectors from the training data of each gender. The MFCC vectors utilized had 12 coefficients, where the 0^{th} coefficient was excluded. To classify a test utterance $X = \{x_1, \ldots, x_T\}$ with the reference centroids $R = \{r_1, \ldots, r_K\}$ from the clustering, the standard average quantization distortion is calculated as in Equation (1), where $d(\cdot, \cdot)$ is the Euclidean distance $\parallel x_t - r_k \parallel$. The smaller the distortion for the utterance, the higher the likelihood for the test utterance X originating from the gender model that holds R.

$$D_q(X, R) = \frac{1}{T} \sum_{t=1}^{T} \min_{1 \leq k \leq K} d(x_t, r_k) \tag{1}$$

2.2. Context-dependent classification

Some previous work has focused on concatenating various speech feature vectors into a single vector for gender classification [10, 7]. However, such a concatenation presents a problem for many unsupervised learning algorithms, as it results in exponentially sparser mappings of observations to a statistical model [11].

The centroid models provided by MFCC clustering give an unlabeled indication of where different units of sound lie in MFCC space. Rather than using these directly in a classifier, we construct Gaussian Mixture Models (GMMs) of the pitch values associated with each MFCC vector that was included in the calculation of the centroid. The motivation of this technique is that the MFCC centroid positions correspond to different *contexts* of sounds, and these contexts can effect the pitch produced. This is evident from various experiments of pitch distributions and ranges for various sounds, and combinations of sounds [12, 13].

The algorithm used for pitch tracking is the one described by Talkin [14], and implemented in the 'Voicebox' toolkit [15].

In the testing stage, the closest centroid to a test utterance MFCC vector is found from both the male and female gender models. The pitch value associated with the test utterance vector is determined as described above, and the likelihood of observing this value from both the male and female pitch GMMs for the chosen MFCC centroids is estimated. Gender is determined by summing the log-likelihoods from the respective male and female pitch GMMs for the observations within an entire spoken utterance.

2.3. Classifier comparison

If both the classifiers described above give the same classification, then there is reasonable confidence that the classification result is correct and the gender is confirmed. However, if there is disagreement, an additional "acoustic loop-back" process is utilized.

2.4. Pitch-shifting loop-back classification

Groen et. al [16] perform a number of experiments related to the human perception of gender in voice for children. Their interest was in investigating the difference in response time between children with high-functioning autism and normal children. The main finding of interest to us is that the response time for gender perception for both groups changes in specific cases, as the pitch of a voice is artificially transformed into subsequent pitch categories by shifting formant ratios and median-pitch levels, from male to female voices. This suggests that the brain process that classifies gender can have different cognitive loads in cases where gender determination is ambiguous. This observation motivates us to propose an extra layer of processing to resolve the classification in cases where the two classifiers disagree, which we take as an indication that the gender information is ambiguous. This processing can be visualized as measuring whether the ambiguous utterance is in fact closer to the male or the female gender in the pattern-space. We do this by small artificial pitch-shifts on the utterance in either the male or female direction, and then re-classifying it with the two classifiers described earlier, to see if they now agree.

The process works as follows: in cases where there is a disagreement in the classification results from the two classifiers, two copies of the utterance are made. Copy 1 is shifted downwards progressively in pitch steps of a semitone, and Copy 2 is shifted upwards progressively in pitch steps of a semitone, to a maximum shift of two semitones. After the first shift, the utterance is re-classified by the two classifiers. If the classifiers now agree on one gender (only), this gender is taken as the class, and the process ends. If not, another shift is applied. Fig. 1 gives an example. With no pitch-shift applied, the two classifiers disagree. Agreement is reached between the classifiers in two situations: either when the pitch is shifted downwards by one semitone, or when shifted upwards by two semitones. Because the utterance requires only one semitone shift downwards to make the classifiers agree on 'male', then this gender is taken as the correct class. The process of upwards/downwards pitch-shifting and reclassification is iterated until one of the following exit conditions is met:

- The classifiers agree on the class 'male' after a downwards pitch-shift, and this shift is smaller than the last upwards pitch-shift, after which they still disagreed. In this case, the gender 'male' is chosen.
- The classifiers agree on the class 'female' after an upwards pitch-shift, and this shift is smaller than the last downwards pitch-shift, after which they still disagreed. In this case, the gender 'female' is chosen.
- The classifiers agree on the class 'male' after a downwards pitch-shift of two semitones, and on the class 'female' after an upwards pitch-shift of two semitones. In this case, the classification made by the acoustic context classifier result is used.
- The pitch has been shifted by two semitones in both directions and the classifiers still disagree. In this case, the classification made by the acoustic context classifier result is used.

Pitch-shifting is done using the 'SoundTouch' audio processing library [17].

3. Results

A number of experiments were performed on the TIMIT [8], ABI-1 [9] and WSJCAM0 [18] corpora. The TIMIT corpus contains 438 male speakers and 192 female speakers, where each speaker speaks 10 phonetically rich short utterances. The ABI-1 corpus subset used contained 145 male speakers and 140 female speakers, where each speaker speaks 3 extracts of 6 seconds each from accent diagnostic passages. The WSJCAM0 corpus subset contained 55 female speakers and 70 male speakers, where each speaker speaks 5 utterances of around 3-5 seconds each. Every experiment, on each corpus, involved first selecting 100 male and 100 female speakers randomly for training models, whilst the rest were used for testing. At the next experiment iteration (under the same conditions) other training/testing sets were randomly chosen. Each experimental condition was tested with 5 different iterations.

Experiments were of 4 training/testing pairs. TIMIT/TIMIT is a classification of TIMIT data based on training over TIMIT data. ABI/ABI is a classification of ABI-1 data based on training over ABI-1 data. TIMIT/ABI is a classification of ABI-1 data based on training over TIMIT data. TIMIT/WSJCAM0 is a classification of WSJCAM0 data based on training over TIMIT data. In TIMIT/ABI and TIMIT/WSJCAM0 experiments, training data was collected from 100 male and 100 female TIMIT speakers, whilst tests were performed on all the ABI-1 and WSJCAM0 speakers. Also we did not pre-process the audio data using normalization techniques such as CMS (cepstral mean subtraction) which are usually performed on identification tasks over mismatched training/testing data. Experiments were performed over different GMM sizes that model pitch distributions for the various MFCC centroids. However, no difference was observed between 4-GMM and 8-GMM models. The results we present are therefore limited to 4-GMM models.

3.1. TIMIT/TIMIT & ABI/ABI performance

The performance results for TIMIT/TIMIT and ABI/ABI tests is shown in Fig 2. The results for TIMIT/TIMIT tests show that the MFCC classifier performance improves gradually as the value of k (number of cluster centroids) increases from 2 to 16. At this point a performance barrier is reached, and no improvement can be seen at higher values of k. However, the context-dependent classification as well as the pitch-shifting loopback classification maintain a steady performance across all values of k. The variance in the results obtained by the context-based classifier and the pitch-shifting loopback classifier we

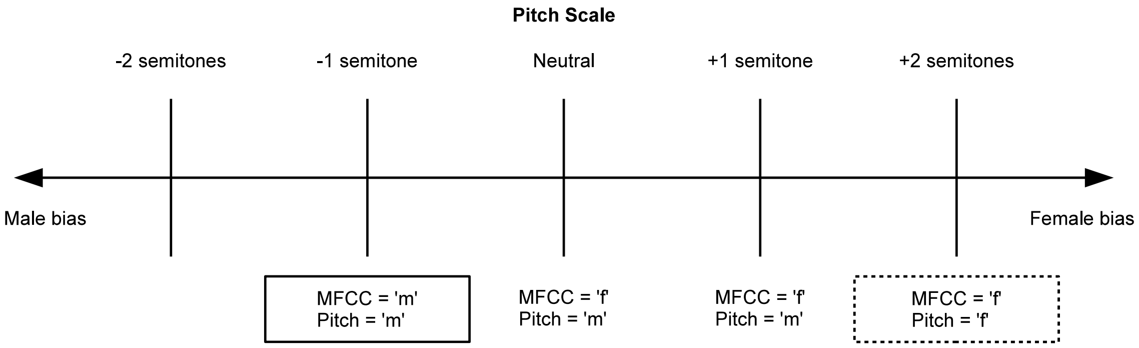

Figure 1: *Pitch-shifting on an utterance. Shifting to the right from the neutral position implies an upward shift of pitch towards the female gender, shifting to the left implies a downwards shift towards the male gender. After each semitone shift, the decisions of the two classifiers ('MFCC' and 'Pitch') are shown. Agreement on the gender is reached after only one semitone shift downwards, but after two semitones upwards, so the utterance is classified as a 'male'.*

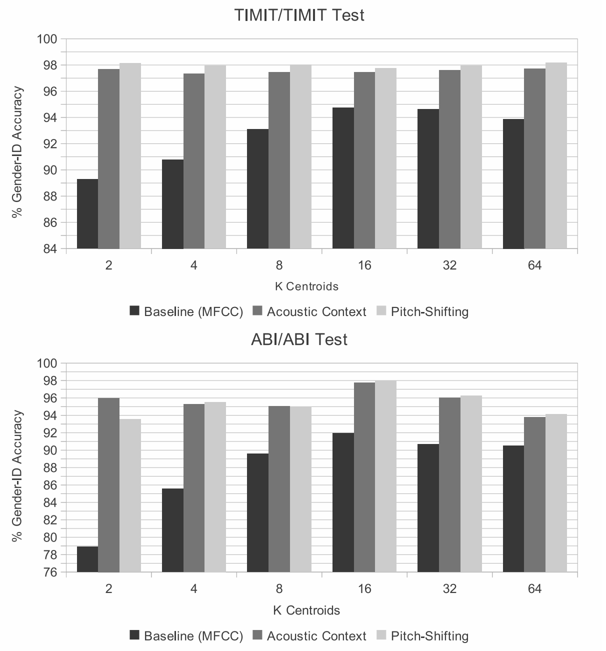

Figure 2: *TIMIT/TIMIT and ABI/ABI test results. The MFCC classifier improves as k increases to 16, and falls off thereafter, whilst context-dependent classification and pitch-shifting are steady across all values of k.*

are proposing in this paper shows greater stability compared to the baseline MFCC classifier, as well as giving higher gender identification performance across all experiments. Also the pitch-shifting loopback classifier always gives a slightly better performance then the context-dependent classifier. The results for ABI/ABI tests show that globally, identification results on all classifiers perform worse on the ABI-1 corpus, when compared to performance on the TIMIT corpus. However both the context-based classifier and the pitch-shifting loopback classifier still perform better than the MFCC classifier. The MFCC classifier performance improves gradually as the value of k (number of cluster centroids) increases from 2 to 16, and drops for $k > 16$. The overall drop in performance on the MFCC classifier (compared to TIMIT/TIMIT tests) is associated with a drop in performance in the other classifiers. Also the pitch-shifting loopback classifier does not always perform better than the context-based classifier, unlike in TIMIT/TIMIT tests.

3.2. TIMIT/ABI & TIMIT/WSJCAM0 performance

The performance results for TIMIT/ABI and TIMIT/WSJCAM0 tests is shown in Fig 3. The results for TIMIT/ABI tests show that the baseline classifier accuracy is lower than in cases where the training and testing sets are the same, and is much lower when training is performed on TIMIT and testing on ABI-1. This indicates that the method is not very robust for classification on different training/testing data sets. The drop in the MFCC classifier is of approximately 30%. On the other hand the context-based classifier maintains a very high and stable classification score in the range of 93-94% accuracy. The pitch-shifting loopback classifier further boosts the results in almost all cases, with a stable result of 95% region for values of $k > 8$, which is very close to the classification accuracy obtained in TIMIT/TIMIT and ABI/ABI tests. The gain for this extra classification stage is therefore greater in TIMIT/ABI tests, and the conclusion is that it is reconciling many errors that occur due to unmatched training/testing data sets. The results for TIMIT/WSJCAM0 tests show that the baseline classifier starts very poorly, in a similar way to TIMIT/ABI tests. The performance improves at higher values of k. However, the performance of the acoustic context and pitch-shifting classifiers is very high on all values of k, further again demonstrating the gain these algorithms have on mismatched training/testing sets.

Table 1: *This table shows how pitch-shifting was utilized across utterances for male and female speakers. Female utterances require the intervention of the pitch-shifting process earlier than male utterances, and in the greater majority of cases require two pitch-shifts.*

	% 0 shifts	% 1 shift	% 2 shifts
Males	65.68%	24.55%	9.77%
Females	46.43%	16.47%	37.10%

The relative number of male and female utterances classified without pitch-shifting and using 1 or 2 semitone shifts is shown in Table 1. Analysis of results indicated that more

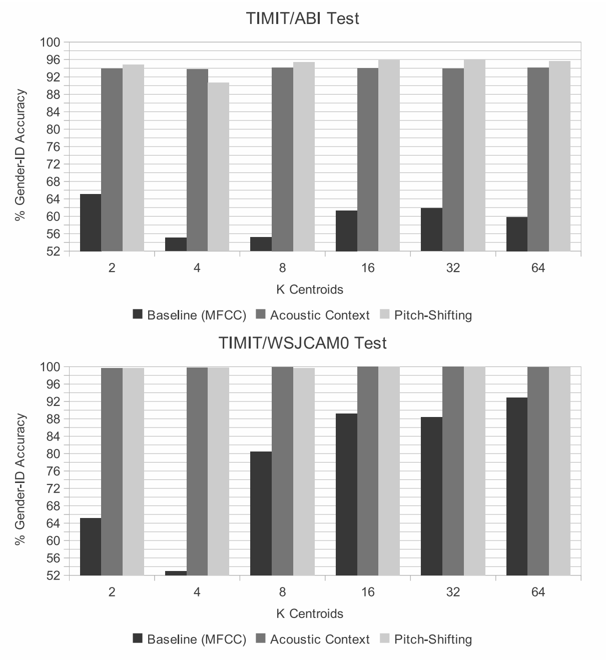

Figure 3: *TIMIT/ABI and TIMIT/WSJCAM0 test results. MFCC classifier improves as k increases, but is very poor compared to context-dependent classification and pitch-shifting which maintain a high and stable result across all experiments.*

than half the utterances from female speakers required a pitch-shifting process before classifiers could agree on gender classification. Also gender-identification on female speech utterances require more intense use of pitch-shifting (2 shifts) than in male speech utterances. This corroborates the experimental results by Groen et. al [16], which concluded that humans take longer to classify gender for female speakers when it is ambiguous than they do for male speakers, and secondly, that more female utterances than male utterances sound ambiguous in pitch.

4. Conclusions

In this paper we have shown how increasing the resolution in specific acoustic regions of speech can be used to build a robust gender classifier. Furthermore, we have described a simple pitch-shifting process guided by classifier fusion, that gives a useful gain in gender identification performance, especially on unmatched training/testing sets. It would be interesting to find other speech features that exhibit similar properties on warping/shifting. In some cases, the upper bound of the accuracy of the MFCC classifier is holding down the potential of the context-based classifier. Therefore a replacement of MFCC features with a feature set that is more gender-specific, rather than speaker-specific, could boost the results of the techniques presented here.

5. Acknowledgments

We would like to thank Mike Brookes (IC, London), for the availability of the 'Voicebox' toolkit, which was used in this work for MFCC and pitch information extraction, clustering and GMM construction. We also thank Olli Parviainen for the availability of the 'SoundTouch' audio processing library, which was used for pitch-shifting. Andrea DeMarco would also like to thank the University of East Anglia for funding his PhD studies.

6. References

[1] K. Wu and D. G. Childers, "Gender recognition from speech. part i: Coarse analysis," *The Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 1828–1840, 1991.

[2] M. Pronobis and M. Magimai.-Doss, "Analysis of f0 and cepstral features for robust automatic gender recognition," Idiap, Idiap-RR Idiap-RR-30-2009, November 2009.

[3] D. Tran and D. Sharma, "Automatic gender recognition," in *Proceedings of the 2nd WSEAS International Conference on Electronics, Control and Signal Processing*. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2003, pp. 49:1–49:5.

[4] Y.-M. Zeng, Z.-Y. Wu, T. Falk, and W.-Y. Chan, "Robust gmm based gender classification using pitch and rasta-plp parameters of speech," in *Machine Learning and Cybernetics, 2006 International Conference on*, 2006, pp. 3376 –3379.

[5] R. Vergin, A. Farhat, and D. O'Shaughnessy, "Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification," in *In Fourth Internalional Conference on Spoken Language Processing*, 1996, pp. 1081–1084.

[6] E. Parris and M. Carey, "Language independent gender identification," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 2, pp. 685–688, 1996.

[7] H. Harb and L. Chen, "Gender identification using a general audio classifier," in *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 1*, ser. ICME '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 733–736.

[8] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," 1993.

[9] Aurix Ltd., "The Accents of the British isles (ABI-1) Speech Corpus." [Online]. Available: http://www.thespeechark.com/abi-1-page.html

[10] J. C. Thankappan and S. M. Idicula, "Language independent voice-based gender identification system," in *Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India*, ser. A2CWiC '10. New York, NY, USA: ACM, 2010, pp. 23:1–23:6.

[11] R. Bellman, *Adaptive control processes: a guided tour*. Princeton University Press, Princeton, N.J.,, 1961.

[12] W. Wei and C. Lianhong, "Research on predicting prosodic parameters for chinese synthesis by data mining approach," *Chinese Journal of Acoustics*, vol. 22, no. 2, pp. 184–192, 2003.

[13] H. Ezzaidi, J. Rouat, and D. O'Shaughnessy, "Towards combining pitch and mfcc for speaker identification systems," in *in Proc. Eurospeech, 2001*, 2001, pp. 2825–2828.

[14] *A Robust Algorithm for Pitch Tracking (RAPT)*. Speech Coding and Synthesis, 1995, ch. 14.

[15] M. Brookes, "VOICEBOX: Speech Processing tool for MATLAB." [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[16] W. Groen, L. van Orsouw, M. Zwiers, S. Swinkels, R. van der Gaag, and J. Buitelaar, "Gender in voice perception in autism," *Journal of Autism and Developmental Disorders*, vol. 38, pp. 1819–1826, 2008.

[17] O. Parviainen, "SoundTouch Audio Processing Library." [Online]. Available: http://www.surina.net/soundtouch/

[18] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," vol. 1, Detroit, 1995, pp. 81–84.