Detection of Children's Voices

Rui Martins[12], *Isabel Trancoso*[12], *Alberto Abad*[2], *Hugo Meinedo*[2]

[1]Intituto Superior Técnico, Lisboa, Portugal
[2]INESC-ID Lisboa, Portugal
rmartins@l2f.inesc-id.pt

Abstract

This paper reports our recent work on extending our previous gender detector, targeted only at distinguishing between adult male and female voices, to encompass children's voices as well. The classifiers were based on multilayer perceptrons and Gaussian mixture models and used Perceptual Linear Prediction coefficients, plus deltas, and pitch as features. Despite the small amount of training data for children's voices, fairly good results were obtained in a test corpus of similar recording conditions (minimum classification error rate of 2.6%). Tests on real life corpora revealed the expected degradation with noisy environments and distant microphones. Tests with transformed female voices intended as cartoon child characters showed that they were mostly classified as children's voices.
Index Terms: gender detection, age effects, children voices.

1. Introduction

Gender detection (GD) is a very useful task for a wide range of applications. In the Spoken Language Systems lab of INESC-ID, the GD module is one of the basic components of our audio segmentation system [1], where it is used prior to speaker clustering, in order to avoid mixing speakers from different genders in the same cluster. Gender information may also be used for building gender-dependent acoustic modules for speech recognition [2]. In our fully automatic Broadcast News subtitling system, deployed at the national TV channel since March 2008 [3], gender information is also used to change the color of the subtitles, thus helping people with hearing difficulties to detect which speaker the subtitle refers to, a useful hint that partially compensates the small latency of the subtitling system.

GD is also a prominent part of our participation in the VIDIVIDEO European project, aiming at the semantic search of audio-visual documents [4]. In this application, the audio concept "male-voice" may be much easier to detect than the corresponding video-concept "male-speaker".

Most gender classification systems are trained for distinguishing between male and female adult voices alone. In fact, in some applications like Broadcast News (BN) transcription, children's voices are relatively rare, hence justifying their non-inclusion. The difficulties in collecting large corpora of children's voices may also be one of the reasons why most detectors do not attempt a 3-class distinction. In some applications such as the automatic detection of child abuse (CA), however, the detection of children's voices may be specially important.

This paper describes our first efforts at moving from our original 2-class gender detection module to a 3-class module including children's voices. The paper starts with a brief overview of the main differences of children's voices relative to adult ones and how they become less pronounced as they grow up. Section 3 reviews the state of the art in terms of features and methods most currently adopted for gender detection. Our own preliminary work is described in 4, before the concluding remarks.

2. Characteristics of children's voices

There are several differences that can distinguish children's voices from adult voices. The differences may be attributed to anatomical and morphological differences in the vocal-tract geometry, less precise control of the articulators and a less refined ability to control suprasegmental aspects as prosody. These aspects induce major differences in children speech, higher fundamental and formant frequencies, greater spectral variability, slower average speaking rate, higher variability in speaking rate and higher degree of spontaneity [5].

It is a well known fact that the fundamental frequency of children's voices is much higher than for adults, where average values of 130 Hz for adult males, and 220Hz for adult females can be found. No statistically significant gender difference exists for children below twelve. Children's voices are also known to have much higher formant frequencies (specially for the second and third formants), attaing values above 4 kHz. The boundary values of the phonetic vowel space decrease with age, becoming more compact, and leading to a decrease in dynamic range of the formants values and to a decrease of the variability of spectral values. A 5-year old child presents values of formants 50% higher than an adult male. Whereas in adults there are typically 3-4 formants in the 0.3-3.2kHz range, for children one can only find 2-3 formants in this range.

2.1. Growing up

The differences become less marked in the process of growing up. During puberty, the male glottis changes so that the pitch frequency is lowered about one octave. This change sometimes occurs over just a couple of weeks. The pitch drop usually occurs from age eleven to age thirteen and there is no significant pitch change after fifteen. No abrupt changes are observed for girls, where the pitch drop from age seven to age twelve is significant, indicating that the laryngeal growth ends around that age. In another study [6] it is shown that for male speakers, pitch drops 78% between the ages 12 to 15, and after that there are no significant changes. For female speakers, pitch drops between ages 7 and 12, and stops after. The changes in female speech are more gradual than in male speech, and the main differences become more significant after age 12.

The size of the vocal tract develops somewhat similarly for boys and girls in this age range [7]. [8] reports an almost linear scaling of formant frequencies with age. The scale presents a significant divergence in male / female after puberty, showing the differences in physical changes between male and female speakers. Another thing that changes with age is the internal

control loops of the articulatory system [9].

3. Gender/age detection

The features most typically found in gender/age classification methods are pitch, formants, Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction Coefficients (PLP), autocorrelation coefficients, linear prediction coefficients (or equivalent), etc. The slower average speaking rate of children relative to adults is also a motivation for including delta, RASTA-PLP, or any other temporal modeling coefficients in the feature set. This large number of features also motivates the adoption of dimensionality reduction approaches such as Independent Component Analysis and Principal Component Analysis [10].

Gender classifiers using Gaussian mixture models (GMM), Hidden Markov models (HMM), or multi-layer perceptrons (MLP) were proposed and tested with results about 95% of accuracy. Most often, these results concern only male/female (M/F) distinction. The comparison of the results reported in the literature is hindered by the fact that they have all been obtained with different corpora. Although very frequently adopted for visual gender detection, Support Vector machines (SVM) are not so popular for audio gender detection.

GMMs are the most frequently adopted learning method for this task. In [11], a two-stage GMM based classifier shows results in the order of 98% accuracy, for clean speech, and Male/Female/Child (M/F/C) distinction, using a feature vector with pitch, formants, and RASTA-PLPs. The first stage attempts to distinguish adult voices from children's voices. The second stage attempts to distinguish between male and female adult voices.

Another GMM-based approach to this problem was proposed in [12], combining the information derived from the pitch with a GMM classifier trained with MFCCs, to enhance the performance of gender classification. The two scores are combined using a weighted summation. This method showed results of 96.7% and 99.7% for sentences and digits, respectively, in an M/F classification task.

Gender male/female detection is also applied in an audio segmentation task for broadcast news in [2]. The authors use an HMM-based phone recognizer with 45 context independent phone models per gender, plus a silence/noise model. The output is a sequence of relatively short segments having male, female or silence tags, which is then heuristically smoothed. The gender segmentation results can be improved by using a clustering procedure in which all segments are clustered using a top-down covariance-based technique. Error rates below 2.4% have been obtained for wideband gender classification.

[14] compares 5 different classifiers for gender and age: multi-layer perceptron, k-nearest-neighbor model, Gaussian mixture model, naive Bayes, and a simple decision tree. Empirical features were adopted: pitch and its microvariations (shimmer and jitter), harmonics-to-noise ratio, articulation rate, number and duration of speech pauses. The age classification was made according to a fine grid: child, teenager, adult, senior. Hence the overall number of classes for the combined gender/age classification problem was 8. The multi-layer perceptron performed best: 93.1% for adult M/F classification, 63.5% for the overall accuracy of the 8-class classification problem. The greatest confusion of the 8-class problem was achieved, as expected in the child M/F distinction.

4. Gender classification experiments

4.1. Corpora

The original Male/Female gender classifier was an MLP trained (and tested) on a corpus of Broadcast News (BN), with approximately 51h, (46h for training and 5h for cross-validation). The first training of the 3-class detector was done using the CMU Kids corpus [15]. The need to get a balanced amount of training data for all the 3 classes made us use a very restricted subset of the BN male/female data (230 min. per class), with corresponding limitations in the classifier results. More recently, however, we had access to the child corpus collected at KTH within the framework of the European project PF-STAR [16]. This allowed us to use an extended corpus around 515 min. per class, which were subdivided into training (345 min.), development (65 min.) and test (105 min.). Table 1 shows the gender / age distribution of the combined corpora.

Gender/Age	4	5	6	7	8	Total
Male	16	36	35	27	18	132
Female	20	27	46	32	18	143

Table 1: Gender/Age Distribution

The two children corpora were recorded in very controlled conditions, and so is most of the BN adult corpus. This was the motivation for building a pilot set of recordings in conditions closer to real-life applications for gender detection. This evaluation corpus includes one broadcast news show (BN - children's day - 63 min.), two TV children's show (CS - 45 min.), two family videos (FV - 30 min.), and 99 CA recordings (489 min.). This CA recordings were divided in 2 sets. The CA Speech which represents word recognizable speech and CA Voice which represents the presence of a human voice (in general with poor acoustic conditions). All have been manually labelled in terms of gender. Results will be presented for each type of show separately, as the conditions widely differ. The BN show is the most similar to the recording conditions of the training corpora - almost no noise, and no speaker overlap. The CS shows are also similar in terms of noise conditions, but multiple speaker overlap is frequent (manually marked as overlapping, with no gender labels). The FV files are characterized by loud background noise and multiple speaker overlap. The CA files often have loud background music.

Very frequently, the voices of child characters in cartoons and games correspond to adult professional speakers. This was the case of the voices chosen for the Ecircus European project [17], where the first recordings of a set of 100 English sentences by a 9-year old girl and a 10-year old boy attested the fact that children have much greater difficulties than adults in recording large quantities of data for corpus-based concatenative synthesis. In fact, they require shorter recording sessions and at slower pace. It is also more difficult to assure the same speaking style among recording sessions, since it often depends on the child mood in that specific day. This was the motivation for building synthetic voices from adult recordings for two female English speakers. The number of prompt files for each speaker was 675. The total duration of the recordings was 24min. for each speaker. Each adult voice was transformed to a child-like voice using PSOLA [18] and spectral scaling techniques. Synthetic voices were then built both from non-transformed and transformed inventories. The transformed voices were considered believable children's voices when played together with the cartoon characters. These 2 pseudo-children's voices will be used

for a last set of tests.

4.2. Classification with GMM and MLP methods

This section reports the experiments with different features and different machine learning methods. The evaluation metrics are the classification error rate (CER), defined as the percentage of incorrectly classified frames, and the F-measure, defined as the weighted harmonic mean of precision and recall.

4.2.1. 2-class baseline classifier

The 2-class baseline classifier [19] is one of the components of the audio segmentation module that is used as a pre-processing stage for our BN fully automatic subtitling system. As other classifiers in this module, it is based on feed-forward fully connected multi-layer perceptrons trained with the back-propagation algorithm. The MLP has 9 input context frames of 26 coefficients (12th order PLP coefficients with energy plus deltas), two hidden layers with 250 sigmoidal units each, and two softmax output units (one for each class) which can be viewed as giving a probabilistic estimate of the input frame belonging to that class.

This classifier was trained and tested with different subsets of the original BN corpus, achieving a CER of 2.30%, with F-measure=0.98.

4.2.2. 3-class MLP classifier

Our first 3-class experiments used an equal MLP architecture, with the same PLP+delta features. As expected, worse results were obtained (CER=4.70%), which we attributed to the drastic reduction in training material, and the addition of a third class. The worse results were obtained for female and children's voices, where the F-measure was 0.95, versus 0.96 for male voices.

Given the importance of pitch as a discriminative feature for this task, we next trained MLPs using PLP+delta+pitch simultaneously, which resulted in an input vector of dimension 27 per frame. Pitch frequency was extracted using the SNACK toolkit [20]. A significant improvement was observed (CER=3.40%). The best results were obtained for male voices, where the F-measure was 0.98, versus 0.96 for female and children's voices.

4.2.3. 3-class GMM classifier

The next set of experiments was done using Gaussian mixture models and the same PLP features plus deltas (26 coefficients). Unlike the described MLP approach, the GMM classifier does not make use of context windows. The number of mixtures was varied from 32 to 512. As expected, best results were achieved for the largest number of mixtures (CER=2.6%). In terms of F-measure, the best results were obtained for children's voices, where the score was 0.99, versus 0.97 for male and 0.96 for female voices.

The next experiment was done using GMMs trained only with pitch information, and varying the number of mixtures from 2 to 32. The results were obviously much worse (minimum CER=30.75%, for 8 mixtures). The highest F-measure was obtained for male voices, where the score was 0.80, versus 0.65 for children's and 0.62 for female voices.

Experiments using both types of features simultaneously yielded CER=4.4% for 512 mixtures. In terms of F-measure, the best results were obtained for male voices, where the score was 0.98, versus 0.96 for children's and 0.94 for female voices.

Results with generative classifiers such as GMMs using both features simultaneously were worse than using only PLP+delta features. It is possible that the fact that pitch values are relatively close for female and children's voices may have a negative influence in the results of generative methods. Discriminative classifiers such as MLPs were not so sensitive to this close proximity.

The final set of GMM experiments combined the PLP-based classifier with the pitch-based classifier. The best combination of weights for the linear classifier was trained using a logistic regression, with the Focal-Multiclass toolkit [21]. The results are slightly worse compared with the previous experiment (CER=5.0%).

The use of 12th order PLP coefficients can be questioned as higher order cepstral coefficients are frequent in speaker identification research. However, our gender classification experiments using 18 PLP coefficients (plus deltas) only showed improvements for adult speakers, at the cost of degrading the results for children voices, making the overall results worse.

5. Results on real-life corpora

Figures 1 and 2 present the results of the MLP and GMM (joint PLP+pitch) classifiers for our different real-life corpora. The corresponding CER results are shown in Table 2, respectively. As expected, the best results were obtained for the BN show, which is the one with the recording conditions closest to the training set.

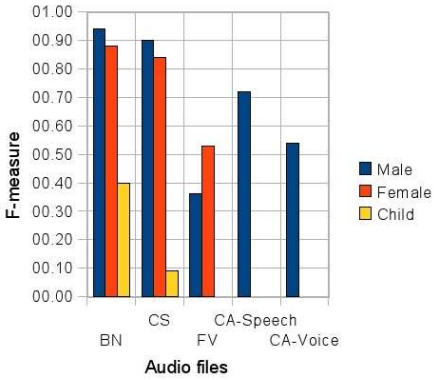

Figure 1: F-measure results on real-life corpora obtained with the MLP classifier (PLP+deltas+pitch).

CER %	BN	CS	FV	CA speech	CA Voice
MLP	10.67	16.64	68.77	44.80	65.27
GMM	18.54	31.96	43.52	47.41	62.55

Table 2: CER results on real-life corpora obtained with the two classifiers (PLP+deltas+pitch).

6. Results on pseudo-children's voices

Experiments with the Ecircus voices have shown us that the original voices were either classified as female or children's voices, which justifies the choice of these particular voices for cartoon child characters. The transformed voices were mostly

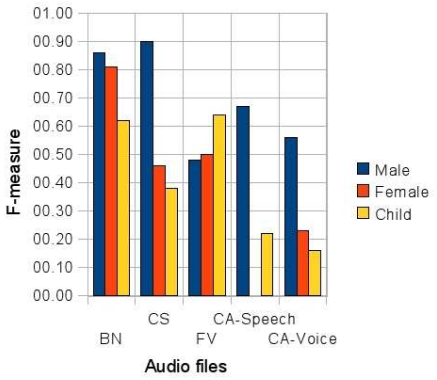

Figure 2: F-measure results on real-life corpora obtained with the GMM classifier (PLP+deltas+pitch).

classified as children's, specially by the GMM classifier, as shown in Table 3.

%		Male	Female	Child
MLP	**Original**	5.38	67.68	26.94
	Transformed	5.55	37.56	56.89
GMM	**Original**	0.23	20.63	79.14
	Transformed	0.11	0.08	99.81

Table 3: Results with the Ecircus voices in terms of percentage of frames classified in each gender/age class.

7. Conclusions

Most of the literature on the detection of children's voices reports results obtained under controlled conditions. It is a well known fact that results generally show a high sensitivity to the presence of noise, and distance to the microphone. Our preliminary experiments with real-life recordings confirm this expected degradation. Nevertheless the results may still be quite useful for a wide range of applications.

The present results show the higher sensitivity of discriminative classifiers when dealing with noisy environments, showing an over-adaption to the clean training environment. Generative classifiers on the other hand are not so accurate in controlled conditions as discriminative ones, but in real conditions they tend to perform significantly better. The fusion of both types of classifier is one of our next tasks.

The reduced amount of training data for children's voices is one of the problems we face. We are currently investigating the possibility of unsupervised training approaches by adding to our children's voices training set all the segments that have been classified as children with a high confidence measure. We are also considering multi-stage classifiers, instead of 3-class ones.

8. Acknowledgements

The authors would like to thank Mats Blomberg and Daniel Elenius for letting us use the KTH PF-STAR children corpus, and our colleague Luís Oliveira for his help with the Ecircus voices. This work is part of the MSc thesis of Rui Martins, and is partly funded by the European projects I-DASH and VIDIVIDEO.

9. References

[1] R. Amaral, H. Meinedo, D. Caseiro, I. Trancoso, and J. Neto, "A prototype system for selective dissemination of broadcast news in european portuguese," *EURASIP Journal on Advances in Signal Processing, Hindawi Publishing Corporation*, no. 37507, May 2007.

[2] P. Woodland, T. Hain, S. Johnson, T. Niesler, A. Tuerk, and S. Young, "Experiments in broadcast news transcription," in *Proc. ICASSP'1998*, Seattle, USA, May 1998.

[3] H. Meinedo, M. Viveiros, and J. Neto, "Evaluation of a live broadcast news subtitling system for portuguese," in *Proc. Interspeech '2008*, Brisbane, Australia, Sep. 2008.

[4] I. Trancoso, T. Pellegrini, J. Portêlo, H. Meinedo, M. Bugalho, A. Abad, and J. Neto, "Audio contributions to semantic video search," in *Proc. ICME 2009 - IEEE International Conf. on Multimedia & Expo*, Cancun, Mexico, 2009.

[5] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in *Proc. International Workshop on Multimedia Signal Processing - MMSP'2007*, Chania, Greece, Oct. 2007.

[6] J. Ajmera, "Effect of age and gender on lp smoothed spectral envelope," in *Proc. Speaker and Language Recognition Workshop, IEEE Odyssey 2006*, San Juan, Puerto Rico, Jun. 2006.

[7] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. ICASSP '1996*, Atlanta, Georgia, USA, May 1996.

[8] S. Potamianos, A.; Narayanan, "Robust recognition of children's speech," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, Nov. 2003.

[9] J. Sundberg, *The Science of the Singing Voice*. Northern Illinois University Press, Dekalb Illinois, 1987.

[10] C. Huang, T. Chen, S. Li, E. Chang, and J. Zhou, "Analysis of speaker variability," in *Proc. Eurospeech '2001*, Aalborg, Denmark, Sep. 2001.

[11] Y. Zeng and Y. Zhang, "Robust children and adults speech classification," in *Fourth Int. Conf. on Fuzzy Systems and Knowledge Discovery - FSKD'2007*, Haikou, China, Aug. 2007, pp. 721–725.

[12] H. Ting, Y. Yingchun, and W. Zhaohui, "Combining mfcc and pitch to enhance the performance of the gender recognition," in *Proc. ICSP 2006*, Guilin, China, Nov. 2006.

[13] K. Hye-Jin, B. Kyungsuk, and Y. Ho-Sub, "Age and gender classification for a home-robot service," in *Proc. 16th IEEE International Conference on Robot and Human Interactive Communication*, Jeju, Korea, May 2007.

[14] C. Mller, "Automatic recognition of speakers' age and gender on the basis of empirical studies," in *Proc. Interspeech '2006*, Pittsburgh, USA, Sep. 2001.

[15] M. Eskenazi, J. Mostow, and D. Graff, "The cmu kids corpus," in *Linguistic Data Consortium*, Philadelphia, USA, 1997.

[16] A. Batliner, M. Blomberg, S. DArcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The pf star childrens speech corpus," in *Proc. Interspeech 2005*, Lisbon, Portugal, Sep. 2005.

[17] C. Weiss, L. Oliveira, S. Paulo, C. Mendes, L. Figueira, M. Vala, P. Sequeira, A. Paiva, T. Vogt, and E. Andre, "Ecircus: Building voices for autonomous speaking agents," in *Proc. 6th ISCA Speech Synthesis Workshop*, Bonn, Germany, Aug. 2007.

[18] F. Charpentier and E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," in *Proc. Eurospeech 1989*, Paris, France, Sep. 1989.

[19] H. Meinedo, "Audio pre-processing and speech recognition for broadcast news," Ph.D. dissertation, Instituto Superior Técnico, Lisbon, Portugal, 2008.

[20] K. Sjolander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proc. ICSLP'2000*, Beijing , China, 2000.

[21] D. van Leeuwen and N. Brummer, "Channel-dependent gmm and multi-class logistic regression," in *Proc. Odyssey'2006*, 2006.