A ZERO-SHOT APPROACH TO IDENTIFYING CHILDREN'S SPEECH IN AUTOMATIC GENDER CLASSIFICATION

Amruta Saraf, Ganesh Sivaraman, Elie Khoury

Pindrop, Atlanta, USA

2022 IEEE Spoken Language Technology Workshop (SLT) | 979-8-3503-9690-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/SLT54892.2023.10023121

ABSTRACT

Detecting whether a speech utterance belongs to an adult male, adult female or a child category, also known as male-female-child (MFC) classification is particularly challenging due to two main reasons - paucity of children's speech data, and high variability in children's speech due to developmental changes. It is difficult to obtain speech datasets with children's voices due to privacy reasons. This paper explores a zero-shot learning approach to MFC classification. Different algorithms are explored to create artificial childlike voices from adult voices. Methods such as pitch shifting, Vocal Tract Length Perturbation, and Segmental Warping Perturbation are used to create synthetic childlike speech for the MFC classification task. Speaker embeddings extracted from a DNN based speaker recognition system are used as features for MFC classification. Compared to a pitch frequency based baseline MFC classifier, the proposed method improves the child classification accuracy by 47%.

***Index Terms*—** Children's speech, gender classification, age estimation, voice attributes, spectral warping, data augmentation, vocal tract length perturbation

1. INTRODUCTION

Identifying speaker characteristics like gender, age, and language from speech is garnering a lot of attention in a wide variety of applications. Gender identification from speech is not limited to male and female classification, but a three-way classification of male, female and child (MFC classification) is gaining importance for access control and personalization of applications. One reason is that children are getting increasing access to voice controlled applications like AI personal assistants or voice controlled television remote control. It becomes especially important to know that the speaker is a child for parental control of AI personal assistant devices and TVs. Moreover, child speech is definitely distinguishable from adult female and adult male speech, so it naturally gives way to the three way classification. In order to guard the privacy of children data, in some applications it may be prohibited to glean more information from the speech once it is known that the speaker is a child.

Children have a short vocal tract undergoing rapid change due to development leading to lower intelligibility of speech compared to adults. In general, children voices tend to be higher in pitch than adult voices. Initial attempts at MFC identification from speech used speech features that captured these age dependent trends, such as pitch [1], and formant frequencies [2]. Later, prosodic and Cepstral features were added for gender classification [3]. In the Interspeech 2010 Paralinguistic Challenge [4], an age and gender classification task was included, in which seven classes, namely child, youth (male and female), adult (male and female) and senior (male and female) were to be identified. Submissions to the challenge proposed additional use of pitch trajectory features [5] and other acoustic frame-based features like MFCCs and Perceptual Linear Prediction (PLP) [6, 7]. Classifiers like SVMs [8], deep neural networks [9] and fuzzy logic [10] have also been explored for the gender classification task. The male, female and child classification is complementary to speaker-age estimation. Age and gender classification outputs are often combined to implement MFC classification. Recent approaches ([11, 12]) have explored the use of speaker embedding vectors from DNNs for the gender classification and/or age estimation. Studies [13, 14] using speaker embedding for adult-vs-child classification are mainly focused on speaker diarization of dyadic interactions. In this paper we use speaker embeddings to classify individual utterances into male, female and child classes. It is worth noting that a dedicated end-to-end MFC classification system might be better using acoustic features such as MFCC, FBank or PLP. However the reason we rely on the speaker embeddings for MFC classification is to have low computational overheads by using an already deployed speaker embedding system which is relevant in industrial settings.

Data scarcity is an obstacle in utilizing DNNs effectively for the MFC classification. We need techniques that can help with data augmentation to generate artificial child like speech data. Performance of ASR systems on children's speech is significantly poor compared to adults due to high variability of children's speech and a scarcity of training data. Data augmentation techniques like applying Vocal Tract Length Perturbation [15] (VTLP) or Segmental Warping Perturbation [16] (SWP) among others have shown to significantly improve the ASR performance on children's speech.

To the best of our knowledge, child speech simulation

978-1-6654-7189-3/22/$31.00 ©2023 IEEE

techniques have not yet been applied for the MFC classification task. In this paper we constrain the problem of MFC classification by avoiding the use of any children's speech data for training. The entire child-class in the training dataset is artificially generated from the adult (female) speech. In Section 2, we describe the child speech simulation techniques that we implement for creating the child-class data for training. We evaluate the MFC classification systems on a test dataset containing real children's speech. In Section 3, we provide the details about the datasets used for training and the test protocol. Thus, our approach to MFC classification is a zero-shot learning approach. Section 4 describes the speaker embedding system which is our front-end feature extractor, and Section 5 explains the MFC classification system. We compare our zero-shot classifier with a pitch thresholding based baseline classifier. All our experiments and results are presented in Section 6.

2. CHILD SPEECH SIMULATION

Children's speech has higher pitch frequency compared to adults. Their pitch frequency ranges from 240 Hz to 300 Hz. Adult female speech exhibits pitch frequencies in the range of 150 Hz to 200 Hz while adult male speech has pitch frequencies in the range of 100 to 150 Hz [17].

Children under the age of 15 have smaller vocal tracts which rapidly grow in length and width as they grow. Their developing control of the vocal tract also leads to poorly articulated speech compared to adults. Due to shorter vocal tract length and higher variability in articulation, for children's speech the formant frequencies (F1 - F4) are higher in magnitude and have higher variance compared to adults [17]. To simulate children's speech from adults, we adopt three methods - 1) Pitch shifting, 2) Vocal Tract Length Perturbation (VTLP), and 3) Segmental Warping Perturbation. We combine phase vocoder based pitch shifting and linear prediction (LP) analysis based spectral warping of formant frequencies (F1-F4) to simulate children's speech. We perform the simulation of children's speech only using female speakers. Since female speech has pitch frequencies close to the range of children's speech, the pitch shift and spectral warping sound more natural when applied to female speech compared to that on male speech.

2.1. Pitch Shifting (pshift)

The pitch of a female speech utterance is shifted by first time stretching the signal using a phase vocoder followed by resampling to match the duration of the input signal. In this paper, we implement the pitch shifting using the librosa tool[1] [18]. Given an input female utterance, we first compute the average pitch of the input utterance, and then randomly choose a pitch shift factor such that the shifted pitch lies in the range of 250 - 300 Hz. Figure 1 shows the block diagram of the random pitch shifting algorithm for simulating child speech.

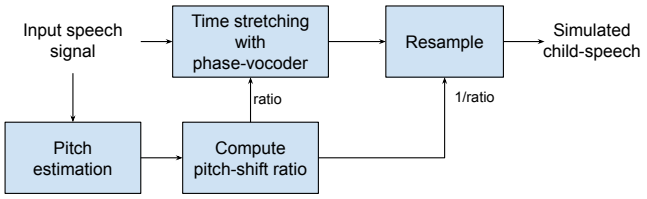

Fig. 1. System diagram of pitch shifting based child-speech simulation.

2.2. Vocal Tract Length Perturbation (VTLP)

In order to simulate the effect of short vocal tracts for children's speech, we use an existing vocal tract length perturbation algorithm [15]. The shorter vocal tracts of children leading to higher formant frequencies (F1-F4). The formant frequencies correspond to the location of the poles in the LP analysis. We implement the VTLP approach in the z-transform domain of the LP filter. We take a simple piecewise linear warping function to upshift the location of the pole frequencies of the input LP spectrum. We then synthesize simulated child speech audio by filtering the original LP residual signal with the warped LP coefficients.

We use an analysis window of 25 ms and a frame shift of 10 ms. We perform LP analysis with an order of 18 on each frame. After applying the VTLP warping on each frame, we reconstruct the audio using LP synthesis and overlap-add.

Equation 1 shows the VTLP frequency warp function. where S is the Sampling frequency and F_{hi} is taken to be 4,800 such that it covers significant formants. f' is the mapped frequency f after the perturbation. We use a random value of α between 0.7 and 0.9 for the VTLP warping. Lower the value of α higher is the up-shift of the formants.

$$f' = \begin{cases} f/\alpha, \text{ if } f \leq F_{hi} \cdot \alpha \cdot \max(1/\alpha, 1) \\ S/2 - \frac{S/2 - F_{hi} \cdot max(1/\alpha, 1)}{S/2 - F_{hi} \cdot \alpha \cdot max(1/\alpha, 1)} \cdot (S/2 - f), \text{otherwise} \end{cases} \quad (1)$$

2.3. Segmental Warping Perturbation (SWP)

SWP is an alternative method of frequency warping to upshift the formant frequencies in the LP spectrum. This method is a recently proposed alternative to VTLP for simulating child speech. In SWP, a different piecewise linear warping is used around each formant frequency to add more non-linearity to the frequency warping. Recent work [16] has shown that the SWP based data augmentation performs better than VTLP for children's speech recognition.

Unlike the VTLP which uses a single warping factor α, SWP uses a different warping factor $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ for the

[1] https://librosa.org/

four segments corresponding to the four formant frequencies. We apply a different warping factor α_i for segments $i = 1, \ldots, 4$. We randomly choose the segmental warping factors $\alpha_1 \in (0.6, 0.85)$, $\alpha_2 \in (0.7, 0.85)$, $\alpha_3 \in (0.75, 0.95)$, $\alpha_4 \in (0.85, 1.0)$ as suggested in [16]. Figure 2 shows a sample LP spectrum warped using SWP. We can see that the segments are defined around each formant frequency and a different warping factor is chosen for each segment. In this paper we limit the segmental warping to the first four formants. No warping is applied to the frames corresponding

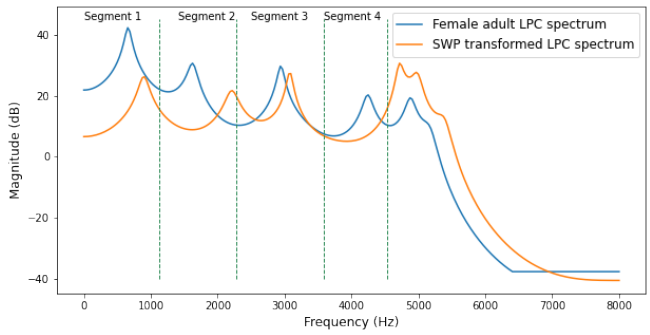

Fig. 2. Example LP spectrum plot showing the segmental warping perturbation.

to unvoiced phonemes where four prominent formants might not be present. The implementation of SWP is similar to the VTLP, the difference only limited to the frequency warping function. The system block diagram for VTLP and SWP are shown in Figure 3.

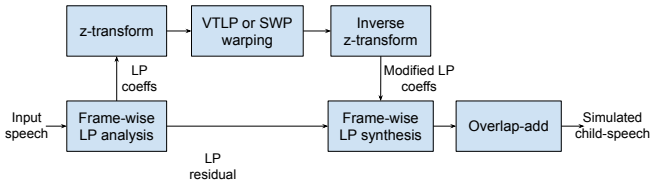

Fig. 3. System diagram of VTLP and SWP based child-speech simulation.

The VTLP and SWP methods only perturb the envelope spectrum of the voiced phonemes in the speech and do not change the fundamental frequencies as we don't modify the LP residual signal before synthesis. In order to increase the pitch of the vocal tract length perturbed speech signal, we apply pitch shifting on the output of VTLP and SWP. In this paper, we explore 3 different approaches for child speech simulation - 1) pitch shifting, 2) VTLP, and 3) SWP and compare the performance of each method in simulating good child-speech training data for male-female-child classification. Note that for both VTLP and SWP techniques, we apply pitch shifting after synthesizing the warped LP spectrum back to time domain.

3. DATASETS

We make use of two datasets to train the zero-shot male-female-child (MFC) classifier - 1) AgeVoxCeleb [19] and 2) Librispeech. The AgeVoxCeleb dataset provides the age labels for a subset of the Voxceleb 2 dataset. We use these to identify child utterances (less than 16 years of age). The train and test protocols of AgeVoxCeleb are used as-is except for one change. We remove the child utterances from train protocol and add them to the test protocol. We create the artificial child samples for the training based on three types of augmentations that we describe in Section 2. We only augment the female data within the training protocol. Thus, we finally have 96,083 male utterances spanning 2,678 speakers, 52,956 female utterances spanning 1,724 speakers and three sets of 52,956 simulated child utterances. We create one copy of the female speech utterances for each child-speech simulation technique described in Section 2. Overall, we end up with 158,868 simulated child-speech utterances.

In the test protocol, which we consistently use across all experiments, we have 10,891 male utterances spanning 340 speakers, 5,637 female utterances spanning 209 speakers and 413 real child utterances spanning 31 child speakers.

We use the Librispeech dataset [20] as an out-of-domain dataset for training the male-female-child classifier. We randomly sample 50 hours of male utterances from the train-clean-100,-360, and train-other-500 subsets spanning 13,600 utterances and 1,206 male speakers. Similarly, we randomly sample around 100 hours of female utterances from the complete Librispeech dataset spanning 27,200 utterances and 1,134 speakers from this dataset. We use half of the female speech dataset (50 hours) for child speech simulation and exclude the corresponding original utterances from the female speech training set. As in the previous dataset, we augment only the female utterances with three types of perturbations to create the copies of simulated child speech data.

We use combinations of the male, female, and simulated child speech utterances from the Voxceleb and the Librispeech datasets for training gender classification systems.

4. SPEAKER EMBEDDING SYSTEM

In this paper, speaker embedding extracted from a DNN based speaker recognition system are the front-end features for the male-female-child classification problem. The speaker embedding system in this paper consists of a fairly simple Convolutional Neural Network (CNN) architecture. An energy-based voice activity detection is used to get rid of silence segments from speech utterances. We then extract 30-dimensional Mel frequency Cepstrum coefficients (MFCC) on 20 ms windows with an overlap of 10 ms. The MFCC features are normalized using zero-mean and unit-variance normalization and then fed into the CNN. The network architecture includes five convolutional layers with 512 filters each,

followed by a statistics pooling layer, two fully-connected layers (with 512 nodes), and a softmax output layer. The output layer consists of 6,114 target speakers. The training is done in two steps. The first step consists of training a speaker embedding system using softmax and categorical cross-entropy. In the second step of training we remove the second fully connected layer and the softmax layer, freeze the remaining layers, and then add a fully connected layer of size 512 x 256 that is trained using large margin cosine loss (LMCL) [21]. The speaker embeddings extracted from the second fully connected layer are used as the feature vector for downstream male-female-child classification. We use the Voxceleb 2 [22] dataset consisting of 6,114 speakers to train the speaker embedding system. The training set for the speaker embedding system does not contain any child speaker.

5. GENDER CLASSIFICATION SYSTEM

The MFC classification system is aimed at classifying an input speech utterance as belonging to one of adult male, adult female or child classes. This three-way classification system which we refer to as MFC classification system in this paper is trained only on adult speech. No real child speech is used to train the classifier, and hence the child class identification is a zero-shot problem in this paper.

5.1. Pitch Based Baseline Classifier

It is well known that the fundamental frequencies of children's speech are much higher compared to that of male and female speech. The adult male pitch ranges from 100-180 Hz, adult female pitch ranges from 180-250 Hz and children's pitch ranges from 250-300 Hz [17]. Our baseline system uses only the average pitch estimated from an input speech utterance as the feature for MFC classification. If the pitch is below 180 Hz, we assign the utterance to the male class. If the pitch is between 180 and 250 Hz, we classify the utterance as female class. All utterances with average pitch estimate greater than 250 Hz are classified as child class. This simple classifier does not need any training. Most widely used adult gender and child classification algorithms are based on pitch and hence our baseline system matches common practice and is also compatible with the challenging zero-shot approach taken in this paper.

5.2. K-Nearest Neighbors (k-NN) Classifier

Our MFC classifier is based on a modified k-nearest neighbors algorithm. We use the speaker embeddings as the input features. The speaker embeddings are widely spread for the male and the female class due to the higher variability in their pitch and vocal tract lengths compared to the child class. The plot shown in figure 4 shows the 2-dimensional t-SNE projections [23] of the speaker embeddings for the male, female, and simulated child speech classes from the Librispeech dataset. We first find C clusters each in the male, female and simulated child class using k-means algorithm. We extract $3C$ the cluster centroids and length normalize them. The C cluster centroids for each class are used as template vectors for the k-nearest neighbor classification. Given an input utterance, we extract the 256 dimensional speaker embedding, and then classify it based on the class identity of the K nearest neighbors in the space consisting of $3C$ class specific centroid vectors. Cosine distance is used as the distance metric for the k-NN classifier. The figure 5 shows the block diagram of the proposed classification system. We chose this modified approach for k-NN classification in order to reduce the effect of outliers affecting the naive k-NN classification. We performed a grid-search on the values of C (range(10,101,10)) and K (range(3,C,2)) with a 10% validation split on the Librispeech training dataset. We found that higher the value of K (close to K = C-1), better are the accuracies of the male and female classes, but lower is the child classification accuracy. Using only top-3 nearest neighbors (K=3) worked best for the child class but poorly for the male and female classes. We chose C=100, and K=51 as a balance between the accuracies of the male, female and child classes. We use $C = 100$, and $K = 51$ for all our experiments in this paper.

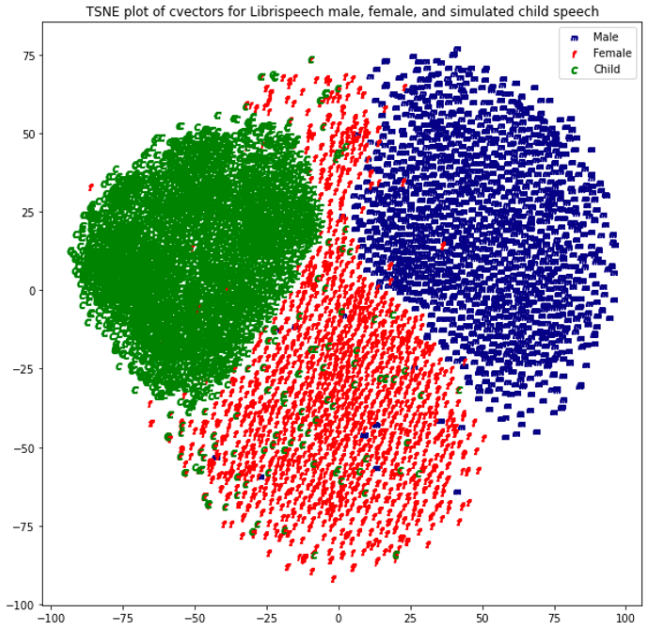

Fig. 4. T-SNE based visualization of the speaker embedding vectors for the Librispeech dataset.

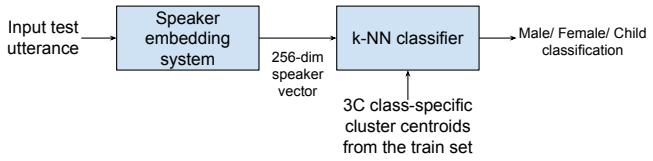

Fig. 5. System diagram of the k-Nearest neighbor based gender classification system.

6. EXPERIMENTS AND RESULTS

All experiments are evaluated on the same Agevoxceleb test subset described in section 3. The pitch thresholding based baseline system is a simple classifier implemented as a threshold on the average pitch frequency of each utterance. We estimate the pitch using the Yin pitch estimator [24] implemented in the librosa tool [18]. We use an off-the-shelf speech enhancement system [25] to reduce the background noise of the age-Voxceleb test utterances before estimating the pitch. The overall baseline system accuracy is 77.06%. It is important to note that the child-class classification accuracy is only 27.54%. Figure 7 shows the confusion matrix of the baseline classifier. There is major confusion between the child class and the female class due the similarity of pitch for children and adult females. The background noise in the data affecting the pitch estimation might also be a major reason for the poor performance of the baseline system.

To further highlight the complexity of the MFC classification task, let us compare the baseline performance of only male vs female (adult) classification on the Agevoxceleb data. The overall classification accuracy of the male vs female (adult) classes on removing the child class is 93.56%, which serves as a kind of upper bound for this simpler task. The male and female class accuracies are 94.50% and 91.73% respectively as compared to 86.8% and 61.9% from Figure 7 of the MFC classification task. Thus, our task's challenge is clear: adding the child class takes a toll on both the adult classes, especially adding confusion to the adult female class.

The pshift, VTLP, and SWP methods simulate children speech with varying audio quality and naturalness. We compare the three methods using the accuracy of MFC classification systems trained on child-speech data simulated by each of these individual methods. The Librispeech dataset which has higher audio quality compared to age-Voxceleb, is more suitable for this comparison. We created three training datasets using the combinations - (libri - m,f,pshift), (libri - m,f,VTLP), (libri - m,f,SWP). In each set, we use the same male-class and female-class utterances from Librispeech. The only variation across the three datasets is the method used to simulate the child-class utterances. Figure 6 shows the bar plot of the test-set accuracies obtained using the three data subsets. We observe that the pshift method outperforms VTLP, and SWP for child-class speech simulation. The performance of the SWP method is significantly poor compared to the VTLP and pshift method. This might be due to the implementation of the SWP warping in the z-transform domain instead of the LP spectrum in the Fourier transform domain.

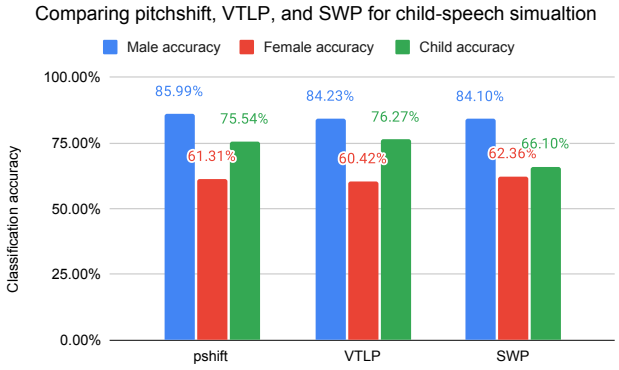

Fig. 6. Comparing child speeech simulation methods with Librispeech data.

Finally, we experimented with different combinations of the age-Voxceleb and the Librispeech data for training the MFC classifier. In all the systems we use only simulated child speech. Table 1 shows the results of the MFC classification using different training data combinations. All the systems outperform the pitch based baseline system. The system trained using ageVoxceleb, and Librispeech data with SWP based child data simulation performs the best with 86.25% overall accuracy. However for this system (agevc+libri, SWP) the child class accuracy is the lowest among the agevc+libri systems. The higher overall accuracy is due to the class imbalance in the test dataset. The highest child-class accuracy is achieved by the "libri, pshift" system at 75.54% which is trained only on Librispeech data. However the overall accuracy of this classifier is only 77.52% due to the difference in domain between the Librispeech data and the AgeVoxceleb test data. The system trained on ageVoxceleb with pshift obtains an overall accuracy of 85.46% which is close to the best performing system at 86.25%. However, adding the Librispeech dataset (real and simulated) to the training reduces the confusion between the female and the child class by 1.61% absolute. Similar to the observation in Figure 6, if we look only at the child-class accuracy, the VTLP and SWP methods perform poorly compared to pshift. The decline in the accuracy is mainly attributed to the increase in the confusion of real child test utterances with the female class. The percentage of child utterances misclassified as female-class for pshift, VTLP, and SWP are 22.28%, 24.21%, and 30.27% respectively. This indicates that the VTLP and SWP based simulated child speech utterances are more unnatural compared to pshift. The VTLP and SWP simulated utterances have much less mixing with adult females. We computed the mean cosine distance between the female and simulated child

class for pshift, VTLP, and SWP. We found that the average female vs child cosine distance for SWP was higher than that for VTLP, and the same for VTLP was higher than that for pshift. At the same time, the average within-child-class cosine distance is lower for VTLP and SWP, implying that these methods end up with more similar sounding voices. Another point to note is that the overall accuracies of the test set are biased towards male class accuracy due to the class imbalance in the test set. We plan to further investigate this phenomenon to propose an improved vocal tract length warping method for child speech simulation.

Table 1. Results of MFC classification with different training data combinations ($agevc$ =ageVoxceleb, $libri$ =Librispeech).

System	Male accuracy	Female accuracy	Child accuracy	Overall accuracy
Baseline	86.78%	61.94%	27.54%	77.06%
libri, pshift	85.99%	61.31%	75.54%	77.52%
agevc, pshift	93.92%	69.90%	74.82%	85.46%
agevc+libri, pshift	94.11%	71.49%	74.58%	86.10%
agevc+libri, VTLP	93.61%	72.72%	72.40%	86.14%
agevc+libri, SWP	92.48%	75.59%	67.31%	86.25%
agevc+libri, pshift+VTLP+SWP	93.71%	72.50%	70.94%	85.84%

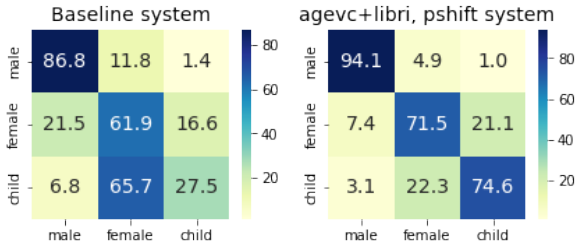

Fig. 7. Confusion matrix of the baseline classifier compared to the best performing system - agevc+libri, pshift.

7. CONCLUSIONS

In this paper, we explored a zero-shot approach to child-speech identification in an automatic gender classification system. We simulated childlike speech from adult female utterances using three different methods - pitch shifting, VTLP, and SWP. Our results show that the pitch shifting based child speech simulation performs better than VTLP and SWP for the MFC classification. VTLP and SWP both warp only the envelope spectrum which is the vocal tract frequency response which mainly affects the articulation of phonemes. Pitch shifting changes the voicing part of the signal which contains more information about the speaker than the content. Hence, pitch shifting might be most effective in simulating childlike speech. We also suspect that the pitch shifting applied after VTLP and SWP may be further shifting the formant frequencies leading to the speech sounding unnatural. We plan to further investigate and improve the VTLP and SWP methods for speaker augmentation.

Our experiments show that child speech can be effectively simulated from adult speech for training an MFC classifier. We can use the speaker embeddings from a system trained only on adult speech as features for MFC classification. If we train the speaker embedding system with simulated child speech as speaker augmentation, it may further improve the child classification accuracy.

The proposed k-NN based classifier is well suited for MFC classification using length normalized speaker embeddings. The proposed system is a minimal-effort classifier without the need for sensitive children's speech data. Compared to the pitch based baseline system, the proposed zero-shot classifier achieves 47% higher child classification accuracy. In the future, we plan to explore a fusion of such zero-shot systems to further improve the performance of MFC classification.

8. REFERENCES

[1] Yu-min Zeng, Zhen-yang Wu, Tiago Falk, and Wai-yip Chan, "Robust gmm based gender classification using pitch and rasta-plp parameters of speech," in *2006 International Conference on Machine Learning and Cybernetics*, 2006, pp. 3376–3379.

[2] 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 *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, 1996, vol. 2, pp. 1081–1084 vol.2.

[3] Ming Li, Kyu J Han, and Shrikanth Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[4] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2794–2797.

[5] Sarah Ita Levitan, Taniya Mishra, and Srinivas Bangalore, "Automatic identification of gender from speech," in *Proceeding of speech prosody*. Semantic Scholar, 2016, pp. 84–88.

[6] Marcel Kockmann, Luk Burget, and Jan ernock, "Brno university of technology system for interspeech 2010 paralinguistic challenge," in *Proc. Interspeech 2010*, 2010, pp. 2822–2825.

[7] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[8] Chih-Chang Chen, Ping-Tsung Lu, Meng-Lin Hsia, Jia-You Ke, and Oscal T.-C. Chen, "Gender-to-age hierarchical recognition for speech," in *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2011, pp. 1–4.

[9] Héctor A Sánchez-Hevia, Roberto Gil-Pita, Manuel Utrilla-Manso, and Manuel Rosa-Zurera, "Age group classification and gender recognition from speech with temporal convolutional neural networks," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 3535–3552, 2022.

[10] Kunjithapatham Meena, Kulumani R Subramaniam, and Muthusamy Gomathy, "Gender classification in speech recognition using fuzzy logic and neural network.," *Int. Arab J. Inf. Technol.*, vol. 10, no. 5, pp. 477–485, 2013.

[11] Damian Kwasny and Daria Hemmerling, "Gender and age estimation methods based on speech using deep neural networks," *Sensors*, vol. 21, no. 14, pp. 4785, 2021.

[12] Nithin Rao Koluguri, Manoj Kumar, So Hyun Kim, Catherine Lord, and Shrikanth Narayanan, "Meta-learning for robust child-adult classification from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8094–8098.

[13] Manoj Kumar, So Hyun Kim, Catherine Lord, and Shrikanth Narayanan, "Improving speaker diarization for naturalistic child-adult conversational interactions using contextual information," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. EL196–EL200, 2020.

[14] Suchitra Krishnamachari, Manoj Kumar, So Hyun Kim, Catherine Lord, and Shrikanth Narayanan, "Developing neural representations for robust child-adult diarization," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 590–597.

[15] Navdeep Jaitly and Geoffrey E Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013, vol. 117, p. 21.

[16] Vishwanath Pratap Singh, Hardik Sailor, Supratik Bhattacharya, and Abhishek Pandey, "Spectral modification based data augmentation for improving end-to-end asr for children's speech," *arXiv preprint arXiv:2203.06600*, 2022.

[17] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, 1999.

[18] Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the 14th Python in Science Conference*, 2015.

[19] Naohiro Tawara, Atsunori Ogawa, Yuki Kitagishi, and Hosana Kamiyama, "Age-vox-celeb: Multi-modal corpus for facial and speech estimation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6963–6967.

[20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. apr 2015, vol. 2015-Augus, pp. 5206–5210, IEEE.

[21] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, dec 2018.

[22] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. sep 2018, pp. 1086–1090, ISCA.

[23] Laurens Van Der Maaten, "Learning a parametric embedding by preserving local structure," in *Journal of Machine Learning Research*, 2009, vol. 5.

[24] Alain de Cheveigné and Hideki Kawahara, "YIN, a fundamental frequency estimator for speech and music.," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[25] Nils L. Westhausen and Bernd T. Meyer, "Dual-signal transformation LSTM network for real-time noise suppression," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020.