Journal Pre-proof

Fundamental frequency feature warping for frequency normalization and data augmentation in child automatic speech recognition

Gary Yeung, Ruchao Fan, Abeer Alwan

PII: S0167-6393(21)00088-1
DOI: https://doi.org/10.1016/j.specom.2021.08.002
Reference: SPECOM 2801

To appear in: *Speech Communication*

Received date : 27 February 2021
Revised date : 27 May 2021
Accepted date : 15 August 2021

Please cite this article as: G. Yeung, R. Fan and A. Alwan, Fundamental frequency feature warping for frequency normalization and data augmentation in child automatic speech recognition. *Speech Communication* (2021), doi: https://doi.org/10.1016/j.specom.2021.08.002.

Fundamental Frequency Feature Warping for Frequency Normalization and Data Augmentation in Child Automatic Speech Recognition

Gary Yeung*, Ruchao Fan, Abeer Alwan

Department of Electrical and Computer Engineering, University of California, Los Angeles, CA 90095, USA

Abstract

Effective child automatic speech recognition (ASR) systems have become increasingly important due to the growing use of interactive technology. Due to the lack of publicly available child speech databases, young child ASR systems often rely on older child or adult speech for training data. However, there is a large acoustic mismatch between child and adult speech. This study proposes a novel fundamental frequency (f_o)-based frequency warping technique for both frequency normalization and data augmentation to combat this acoustic mismatch and address the lack of available child speech training data. The technique is inspired by the tonotopic distances between formants and f_o, developed to model human vowel perception. The tonotopic distances are reformulated as a linear relationship between f_o and vowel formants on the Mel scale. This reformulation is verified using f_o and formant measurements from child utterances. The relationship is further generalized such that the frequency warping technique only relies on two parameters. The LibriSpeech ASR corpus is used for training, and both the OGI Kids' Speech and CMU Kids Corpora are used for both training and testing. A single word ASR experiment and a continuous read speech ASR experiment are performed to evaluate the f_o-based frequency normalization and data augmentation techniques. In the single word experiment, the system using f_o-based frequency normalization significantly improved over the baseline system with no normalization, with a relative improvement of up to 22.3%, when the mismatch between training and testing data was large. In the continuous speech experiment, the combination of f_o-based frequency normalization and data augmentation resulted in a relative improvement of 19.3% over the baseline. Additionally, in all experiments, the f_o-based techniques outperformed other techniques such as vocal tract length normalization (VTLN) or vocal tract length perturbation (VTLP). Results were validated using Gaussian mixture model (GMM), deep neural network (DNN), and bidirectional long-short term memory (BLSTM) acoustic models.

Keywords: child speech, speech recognition, frequency normalization, data augmentation, fundamental frequency

1. Introduction

The need for child automatic speech recognition (ASR) has grown dramatically in recent years. A major reason for this is the increased usage of electronic home devices and living-room personal assistants. Often, speech is one of the only mechanisms young children have to interact with such devices due to their limited reading, writing, and typing abilities. Furthermore, improved child ASR performance can greatly benefit the development of teaching, assessment, and clinical diagnostic tools (Kewley-Port et al., 1991; Tepperman et al., 2006; Bunnell et al., 2000; Yeung et al., 2017; Sadeghian and Zahorian, 2015) through interactive media such as social robots (Kennedy et al., 2017; Spaulding et al., 2018; Yeung et al., 2019). Yet, while adult ASR has experienced significant performance improvement in recent years, child ASR continues to perform quite poorly in comparison (Kennedy et al., 2017; Gerosa et al., 2009).

Previous analyses of child ASR have revealed that the current performance is inadequate for practical usage. For instance, Kennedy et al. (2017) examined the ASR performance of 5 year old child speech using the Alderbaran NAO, a social robot commonly used for human robot interaction (HRI) research. In that study, the ASR system performed insufficiently on even the most basic tasks. This included digit recognition, which had a word error rate of over 15%, and scripted speech recognition, which had a sentence error rate of over 88% on four commercial ASR APIs (Google, Bing, Sphinx, Nuance).

A significant impediment to the development of child ASR is the lack of publicly available child speech databases, especially for young child speech. This is further complicated by the fact that deep learning, which requires large amounts of speech data to train, is becoming the most prominent method of developing ASR systems. To compensate for this lack of data, young child ASR systems often employ speech data from other domains, such as older child speech or even adult speech, to supplement the training data. However, there are many differences between child and adult speech acoustics, further complicated by the fact that children's speech acoustics change as they grow (Lee et al., 1997, 1999; Vorperian and Kent, 2007; Smith, 1992; Koenig et al., 2008; Koenig and Lucero, 2008).

☆Parts of this article appeared in the proceedings of INTERSPEECH 2019 and will appear in the proceedings of ICASSP 2021.

*Corresponding author

Email addresses: garyyeung@g.ucla.edu (Gary Yeung), fanruchao@g.ucla.edu (Ruchao Fan), alwan@ee.ucla.edu (Abeer Alwan)

These changes include the rapid lowering of the fundamental frequency (f_o) and formant frequencies (Lee et al., 1997, 1999; Vorperian and Kent, 2007), two defining acoustic features of the speech space.

Previous studies have investigated the performance of child ASR systems on various age ranges and groupings. Shivakumar et al. (2014) examined child ASR systems using Gaussian mixture model (GMM) hidden Markov model (HMM)-based acoustic models and found that small differences in a child's age can result in dramatic performance changes. Similarly, Yeung and Alwan (2018) examined child ASR systems using both GMM-HMM-based and deep neural network (DNN)-HMM-based acoustic models and also discovered that the age of the speakers can have a significant effect on system performance. In particular, the ASR performance for kindergarten speakers resulted in significantly worse performance than for children just one year older.

Several techniques have been proposed to reduce this mismatch in ASR systems. Frequency normalization techniques attempt to warp the speech spectra to a normalized speech space, reducing inter-speaker variability in the spectral domain. For instance, vocal tract length normalization (VTLN) (Lee and Rose, 1998) uses a maximum likelihood approach to warping the speech spectra. Various implementations of these warping techniques have seen success in child ASR (Kathania et al., 2020; Shivakumar et al., 2014; Stemmer et al., 2003; Cui and Alwan, 2005; Serizel and Giuliani, 2014; Panchapagesan and Alwan, 2006; Gray et al., 2014). Alternatively, acoustically relevant speech parameters, such as the subglottal resonances (SGRs) (Guo et al., 2015) or third formant frequency (F3) (Cui and Alwan, 2006), can be used as normalization factors by warping the spectra to match a default speaker.

More recently, ASR systems based on deep learning have used data augmentation techniques to increase the available data to train large neural networks such as bidirectional long-short term memory (BLSTM) networks. There are several ways to implement these techniques such as feature warping (Jaitly and Hinton, 2013; Cui et al., 2014), adding noise (Ko et al., 2017; Hannun et al., 2014), and masking in time or frequency (Park et al., 2019). An analogue to VTLN, vocal tract length perturbation (VTLP) uses VTLN warping factors to extract features from the same utterance several times, creating additional variability in the training data. Furthermore, data augmentation has not yet been fully explored for child speech, but some known techniques include adding noise and reverberation (Wu et al., 2019) and applying out-of-domain adult speech to the training data (Fainberg et al., 2016; Sheng et al., 2019). Notably, while data augmentation techniques increase the amount of available training data, many techniques simply create variability without considering whether these additional features adhere to the acoustic properties of speech.

Another common technique used in training child ASR systems is system retraining and transfer learning. For DNN-based child ASR systems, one common approach is to first train the system on more readily available data such as adult speech data before fine-tuning the network parameters with child speech data (Wu, 2020; Shivakumar and Georgiou, 2020). Transfer learning from adult data has also seen success in combination with normalization techniques (Shivakumar and Georgiou, 2020).

The use of f_o as a speech feature has also been used successfully in adult ASR. Several past studies have reported that the inclusion of f_o or some voicing parameter as an input feature improved ASR performance for adults, even in atonal languages (Ghahremani et al., 2014; Fujinaga et al., 2001; Ljolje, 2002; Magimai-Doss et al., 2003). Faria and Gelbart (2005) found that f_o could be used to predict the VTLN warping factor of an utterance with a maximum likelihood approach. Shahnawazuddin et al. (2016) used f_o to determine lifter sizes when extracting cepstral features. Furthermore, while many studies examining the use of f_o in ASR were performed on adults, f_o may also be relevant to children. That is, considering the differences in the f_o values of adults and children, f_o is likely to contain relevant speaker information.

Research on human speech perception can provide further insight into the use of f_o in ASR. The tonotopic distances between formants, the distance between adjacent formants in some perceptual scale such as Mel or Bark, along with the tonotopic distance between the first formant and f_o, are a set of features that have been successfully used to model human vowel perception (Chistovich and Lublinskaya, 1979; Chistovich, 1985; Syrdal and Gopal, 1986; Traunmüller, 1981; Fahey et al., 1996). This set of features can be interpreted as a normalization of formant-based vowel models. The inclusion of f_o suggests that f_o contains information that can be exploited to normalize the vowel spectrum. This is supported by studies that suggest that the perception of vowel quality, vowel production, and voice naturalness are dependent on f_o when formants are fixed (Barreda and Nearey, 2012, 2013; Assmann and Nearey, 2007). Furthermore, f_o and the tonotopic distances may also be useful for creating speech features that are more speech-like than other data augmentation methods.

This study proposes an f_o-based feature warping method, inspired by the role of f_o in vowel perception, for both frequency normalization and data augmentation in child ASR. We demonstrate that the tonotopic distances between formants and f_o can be reformulated as a frequency warping function dependent solely on f_o. First, the f_o-based frequency normalization is examined using a data limited single-word experiment against unwarped features and other normalization procedures. Then, both the f_o-based frequency normalization and data augmentation techniques are examined using a continuous read speech experiment with transfer learning from adult speech. In both experiments, the f_o-based methods performed the best.

The remainder of the paper is organized as follows. Section 2 discusses the formulation of the f_o-based warping method for both frequency normalization and data augmentation. Section 3 describes the databases and experimental setups used for the single-word and continuous speech experiments. Section 4 discusses the results of the experiments. Section 5 concludes the paper with a summary and directions for future work.

2. Methods and Formulation

2.1. Tonotopic Distances in Vowel Perception

A number of previous studies has found success in modeling human vowel perception using the tonotopic distances between adjacent formants ($F(n+1) - F(n)$ for $n \in \{1, 2, 3, \ldots\}$), along with the tonotopic distance between the first formant and fundamental frequency ($F1 - f_o$), in the Mel or Bark scale (Chistovich and Lublinskaya, 1979; Chistovich, 1985; Syrdal and Gopal, 1986; Traunmüller, 1981; Fahey et al., 1996). An equivalent representation of this set of tonotopic distances can be formulated as the difference between f_o and each formant ($F(n) - f_o$ for $n \in \{1, 2, 3, \ldots\}$) (Yeung and Alwan, 2019). This reformulation contains several implications for vowel space modeling. Specifically, this implies that a linear relationship (with a slope of 1) exists between a vowel's formants and f_o in the perceptual frequency scale. It follows that the perception of formant locations of vowels is affected by the value of the corresponding f_o. Additionally, this suggests that humans are capable of using f_o to perceive the vowel quality of an utterance. These implications are particularly important for child speech as they have much higher formant and f_o values than adults.

2.2. Tonotopic Distances in Child Vowels

To verify this relationship between formants and f_o, we analyzed utterances of hVd words from the WashU-UCLA Child Subglottal Resonances Database (Yeung et al., 2018). This database includes 43 children, between the ages of 6 and 18 years, saying 14 different hVd words in the carrier phrase, "I said a hVd again." Each hVd word was repeated at least 6 times by each of the children. While this database includes both microphone and subglottal accelerometer recordings for each utterance, only microphone recordings were used for this analysis.

Of the 14 hVd words, four tense vowels, /i/, /æ/, /ɑ/ and /u/, and four lax vowels, /ɪ/, /ɛ/, /ʌ/, and /ʊ/, were chosen for analysis. The values of f_o, $F1$, $F2$, and $F3$ were measured from the vowels of these hVd words. All measurements were done using Praat with manual corrections as needed. Least-squares linear regression lines relating $F1$, $F2$, and $F3$ to f_o for each vowel were computed, resulting in 24 regression lines. Of all the linear regressions, the slopes from 19 of the 24 regression lines were between 0.70 and 1.30, reasonably close to the expected slope of 1. These slopes contributed significantly to the regression ($p < 0.001$) with Pearson's correlation coefficients greater than $r > 0.5$. These results are consistent with the reformulation of the tonotopic distances.

The formant and f_o data of the 8 vowels are displayed in Figure 1 along with least-squares regression lines fixed to have a slope of 1. The data clearly follow the regression lines with an upward trend in formant values as f_o increases. This result demonstrates the validity of the reformulation. It should be noted that as f_o increases, the variability of the individual formants for some of the vowels also increases such as for $F1$ of the vowel /i/. This is mostly attributed to mispronunciations, speech variability, and developing motor skills of children. However, even with this variability, the linear trend still remains.

2.3. f_o-based Normalization

Based on the tonotopic distance reformulation above, we can derive a frequency normalization method in the spectral domain that relies solely on the value of f_o. The key to this derivation is to note that when f_o and the formants ($F1$, $F2$, $F3$, ...) are measured on a perceptual scale, the difference between formants and f_o ($F1 - f_o$, $F2 - f_o$, $F3 - f_o$, ...) should be constant across different productions of the same vowel. Thus, if we can measure f_o for any vowel utterance, we can normalize the formants of the vowel to some default values as follows:

$$F(n)_{norm} = F(n)_{orig} - (f_{o,utt} - f_{o,def}) \tag{1}$$

for $n \in \{1, 2, 3, \ldots\}$ where $f_{o,utt}$ is the f_o of the utterance, $f_{o,def}$ is a predetermined value of f_o to represent a default speaker, $F(n)_{orig}$ is the n-th formant in the original utterance, $F(n)_{norm}$ is the n-th formant after normalizing to $f_{o,def}$, and all frequencies in Eq. 1 are measured in the perceptual scale. This paper will use the Mel scale as the perceptual scale of choice.

While Eq. 1 is formulated specifically for formants, this comes with the added complication of formant estimation, which is generally unreliable for children or speakers with high f_o and formant values. Instead, we can avoid direct manipulation of formants by normalizing the entire spectrum as follows:

$$f_{norm} = f_{orig} - (f_{o,utt} - f_{o,def}) \tag{2}$$

where f_{orig} is some frequency in the original spectrum and f_{norm} is the corresponding frequency in the normalized spectrum. That is, the frequency content in f_{orig} is shifted to f_{norm}. In the case of a discrete spectrum such as for feature computation, we can reinterpret f_{norm} as the normalized frequency corresponding to some discrete Fourier transform (DFT) index and f_{orig} as the frequency from the original spectrum mapped to the index of f_{norm}. We will refer to this method as f_o normalization.

The difference between f_o normalization and VTLN should be noted. While each technique performs some warping of the frequency space, the warping function differs. Specifically, VTLN and variations of VTLN that use physical parameters such as F3 normalization (Cui and Alwan, 2006) and SGR normalization (Guo et al., 2015) are generally implemented using piecewise-linear functions. However, f_o normalization is nonlinear due to the shift in the Mel scale rather than Hz. Additionally, while VTLN usually requires several feature extractions and passes through the ASR system to be effective, f_o, F3, and SGR normalization all use physical parameters to compute the warping function and thus only require a single feature extraction.

An example of the effect of f_o normalization is shown in Figure 2. The Mel filter bank outputs of an 18 year old male and 7 year old male saying the vowel /i/ are displayed both with and without f_o normalization. As expected, when f_o normalization is applied, the filter bank outputs become more similar.

The values of $F2$ vs. $F1$ for each of the utterances of the vowels /i/, /æ/, /ɑ/, and /u/ from the microphone signals of

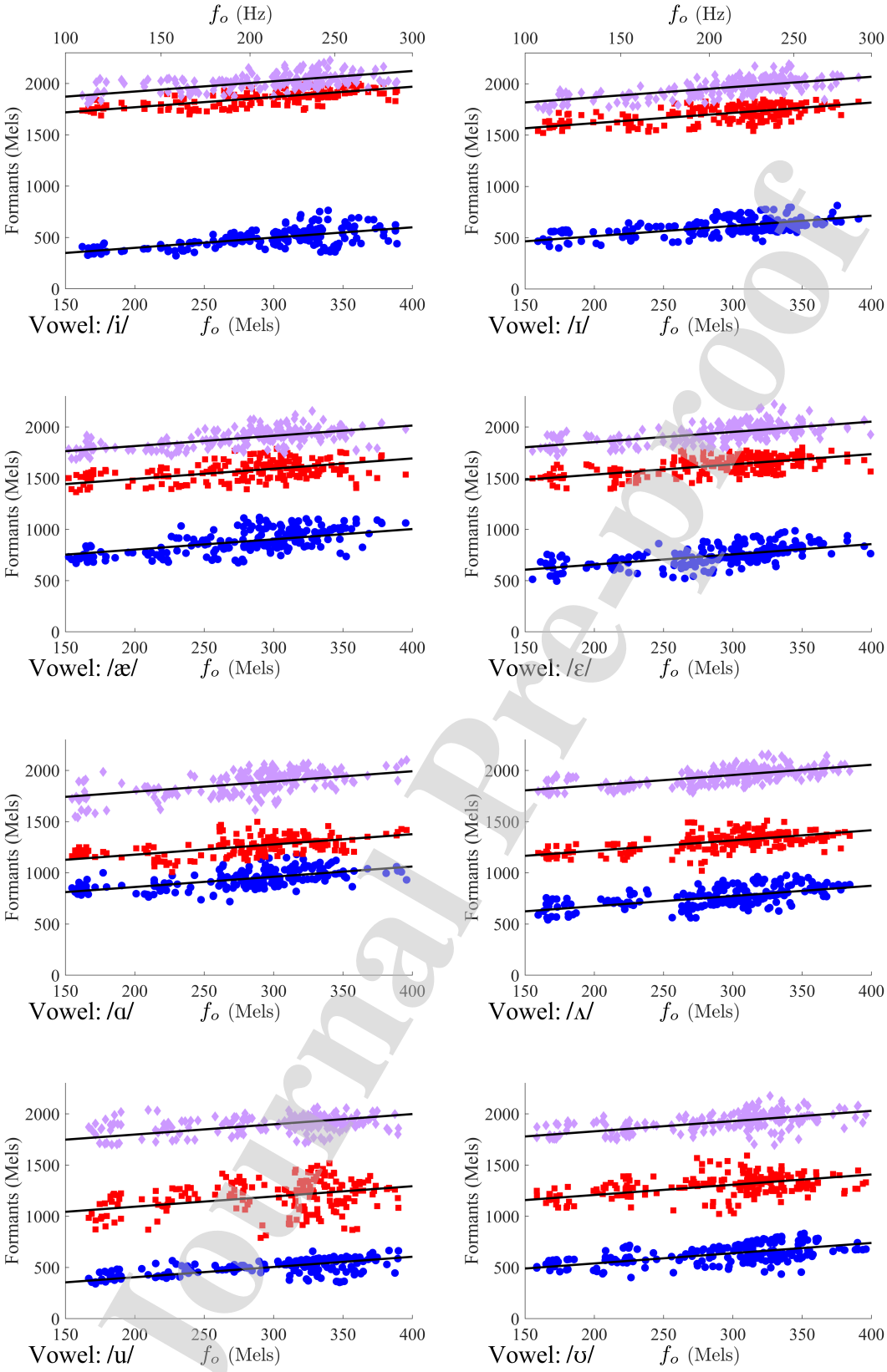

Figure 1: $F1$ vs. f_o (blue), $F2$ vs. f_o (red), and $F3$ vs. f_o (purple) for the vowels /i/, /æ/, /ɑ/,/u/, /ɪ/, /ɛ/, /ʌ/, and /ʊ/ from corresponding hVd words of children between the ages of 6 and 18 years. Also shown are the least-squares linear regression lines, fixed to have a slope of 1. The data follow the linear relationship implied by the reformulation of the tonotopic distances.

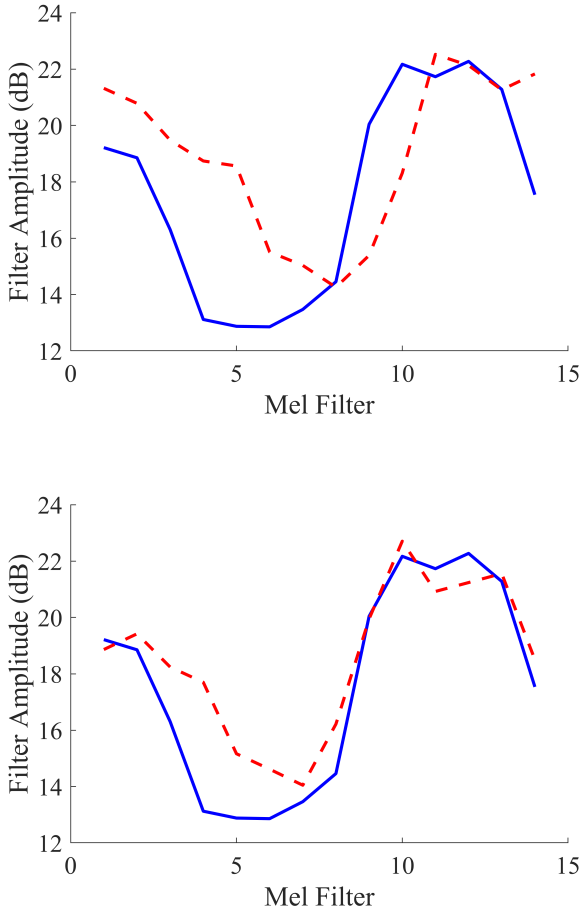

Figure 2: Mel filter bank outputs of the vowel /i/ spoken by an 18 year old male (solid) and a 7 year old male (dashed) computed with 15 filter banks using a frequency range of 20 Hz to 6 kHz. The filter bank outputs are computed both without (top) and with (bottom) f_o normalization. When normalization is applied, default f_o is chosen to be $f_{o,def}$ = 100 Hz. The 18 year old male had $f_{o,utt}$ = 106 Hz, and the 7 year old male had $f_{o,utt}$ = 270 Hz. The filter bank outputs computed with f_o normalization are better aligned than the outputs without f_o normalization.

the WashU-UCLA Child Subglottal Resonances Database are shown in Figure 3. The top figure shows the formants without normalization, and the bottom figure shows the formants with f_o normalization and $f_{o,def}$ = 100 Hz. Compared to the formants without normalization, the $F1$ and $F2$ values for each vowel are clearly more condensed and better separated from the other vowels. This further exemplifies the objective of f_o normalization.

2.4. Data Augmentation via f_o Perturbation

While the f_o normalization procedure uses Eq. 2 to reduce variability between speakers by fixing $f_{o,def}$ to a default value and adjusting $f_{o,utt}$, an alternative procedure is to use Eq. 2 to create variability. To perform this procedure, we extract features several times from a single speech utterance while changing $f_{o,def}$. The resulting set of features is consistent with the

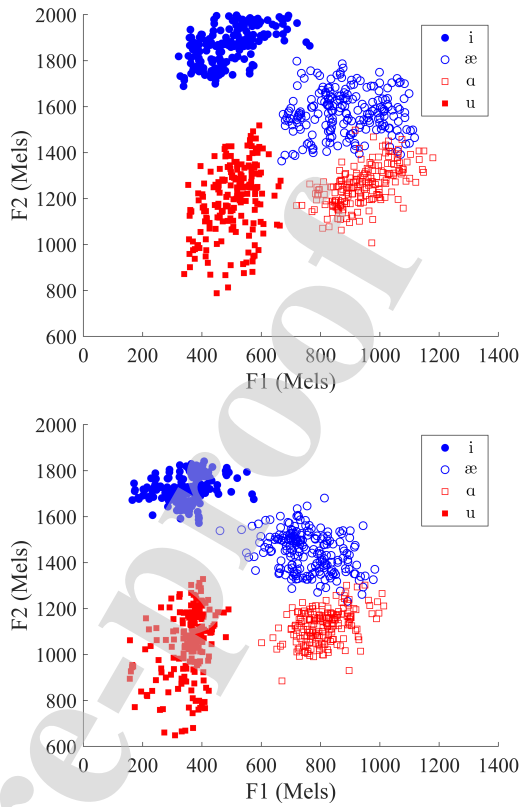

Figure 3: $F2$ vs. $F1$ for the vowels /i/, /æ/, /ɑ/, and /u/, both without (top) and with (bottom) f_o normalization, from corresponding hVd words of children between the ages of 6 and 18 years. When f_o normalization is applied, the default f_o is chosen to be $f_{o,def}$ = 100 Hz. Clearly, the ($F1$, $F2$) locations for the formants with f_o normalization are more condensed and better separated than for the formants without normalization.

structure of speech defined by the tonotopic distances and can be used to augment the training data. This is especially useful for training neural network-based systems, which require large amounts of training data. We will refer to this method as f_o perturbation.

Notably, f_o normalization and f_o perturbation can be used simultaneously by setting $f_{o,utt}$ to be the f_o of the utterance and choosing multiple values for $f_{o,def}$. This allows us to remove large inter-speaker variabilities while also generating speech-like features for additional training data.

2.5. Parameter Considerations

For f_o normalization to be effective, the default value of $f_{o,def}$ must be chosen beforehand so features can be normalized to the same speech space. After preliminary experimentation, we have found that a reasonable choice for $f_{o,def}$ is 100 Hz or 150 Mels. This can be interpreted as normalizing a speaker's voice to that of an adult male. The following experiments in Section 3 will use $f_{o,def}$ = 100 Hz as an initial value.

To perform f_o perturbation, additional values for $f_{o,def}$ must be chosen. This can be done by simply perturbing the initially chosen value of $f_{o,def}$. After preliminary experimentation, we have found that adding ±20, ±40, and ±60 Mels to the initial

value of $f_{o,def}$ is effective. For instance, when choosing an initial value of $f_{o,def}$ = 100 Hz, the f_o perturbation method would use $f_{o,def} \in \{58.52, 72.10, 85.93, 100.00, 114.32, 128.90, 143.74\}$ Hz, repeating the feature extraction for each value.

An additional consideration is the computation of $f_{o,utt}$ required to perform f_o normalization. Generally, any reasonably accurate pitch estimation method can be used for this normalization. Any algorithm that estimates f_o can implement f_o normalization with no computational overhead. The experiments described in Section 3 use the Multi-Band Summary Correlogram (MBSC) pitch detection algorithm (Tan and Alwan, 2013) to compute $f_{o,utt}$.

It is desirable to use the median f_o over the speech utterance as the value of $f_{o,utt}$. Using a median can remove errors and outliers from the f_o estimation process and compensate for intra-speaker variability. Storage of a single number to be applied to an individual speaker is also an effective way to quickly adapt an ASR system as opposed to needing multiple values of f_o for an individual speaker. Finally, DFT computation over an utterance can be implemented much more efficiently when applying f_o normalization with only one value of $f_{o,utt}$ as opposed to many, resulting in dramatically faster computation times for feature extraction. The following experiments in Section 3 will use median f_o over the utterance as the value of $f_{o,utt}$.

2.6. Bandwidth Considerations

When applying the f_o normalization method to feature extraction, it is necessary to consider the usable bandwidth of the speech signal. This is because the normalization method shifts the entire dynamic range of the frequency spectrum. For example, consider an f_o normalization setup with $f_{o,def}$ = 100 Hz and a feature extraction procedure using frequencies from 20 Hz to 8 kHz, which is standard for speech signals sampled at 16 kHz. If an utterance has $f_{o,utt}$ = 100 Hz, the frequency range of the feature extraction does not change. However, if an utterance has $f_{o,utt}$ = 200 Hz, the range of frequencies is shifted by approximately 133 Mels. Thus, the lowest needed frequency shifts from 20 Hz to 110 Hz, and the highest needed frequency shifts from 8 kHz to 9.1 kHz. Similarly, if an utterance has $f_{o,utt}$ = 300 Hz, the range of frequencies used is from 200 Hz to 10.2 kHz. As such, some values of $f_{o,utt}$ would require frequencies that are beyond the signal bandwidth.

To compensate for this shift, the feature extraction bandwidth at $f_{o,def} = f_{o,utt}$ must be limited such that the use of f_o warping does not exceed the signal's bandwidth. In the case of $f_{o,def}$ = 100 Hz, limiting the default bandwidth to 20-6200 Hz ensures that the feature extraction will not use frequencies above 8 kHz for a maximum $f_{o,utt}$ of 300 Hz. This guarantee is sufficient for most child utterances sampled at 16 kHz, and experiments in Section 3 will use a bandwidth of 20-6200 Hz when applying f_o normalization during feature extraction.

3. Experiments

3.1. Databases

Several databases containing both adult and child speech were used in this study. For adult speech data, the LibriSpeech ASR corpus was used (Panayotov et al., 2015). This corpus contains adult read speech (1210 males, 1128 females) from various audio books. The speech in this corpus was sampled at 16 kHz. We specifically used the training set from this corpus, which contains 960 hours of clean and noisy speech.

Table 1: Number of subjects and utterances in the OGI Kids' Speech Corpus, separated by grade level and gender.

Grade	Male		Female	
	# Sub.	# Utt.	# Sub.	# Utt.
K	39	1142	49	1915
1	58	3921	31	2032
2	53	3584	61	2032
3	63	4194	52	3516
4	47	3178	45	2976
5	49	3361	49	3362
6	57	3912	55	3774
7	46	3136	51	3499
8	49	3362	50	3431
9	69	4606	40	2677
10	75	5084	29	1989

Table 2: Number of subjects and utterances in the CMU Kids Corpus, separated by grade level and gender.

Grade	Male		Female	
	# Sub.	# Utt.	# Sub.	# Utt.
K	0	0	1	1
1	6	176	16	944
2	10	812	22	1388
3	8	758	11	942
6	0	0	1	93
Unknown	0	0	1	66

The first child speech database was the OGI Kids' Speech Corpus (also know as the CSLU Kids' Speech Corpus) (Shobaki et al., 2000). This corpus contains read and spontaneous speech from approximately 100 speakers at each educational grade level, from kindergarten to 10th grade. The number of male and female children from each grade level is shown in Table 1. Utterances were sampled at 16 kHz. In this study, we only used read speech utterances that were labeled "1" in the corpus verification files. This label indicates that the utterance was read accurately with limited background noise. The read speech was further split into two datasets depending on whether the child read only a single word or a full sentence. We will refer to these datasets as "OGI single words" and "OGI sentences". Note that the OGI sentences dataset only contained speech from kindergarten to 5th grade.

The second child speech database was the CMU Kids Corpus (Eskenazi et al., 1997). This corpus contains read speech from 76 children between 1st and 3rd grade with two additional speakers from kindergarten and 6th grade. The number of male and female children from each grade level is shown in Table 2. Utterances were sampled at 16 kHz. Children were instructed

to read several short sentences resulting in a total of 5180 sentence utterances.

3.2. Feature Extraction

The baseline features for the following experiments were Mel-frequency cepstral coefficients (MFCCs). The MFCCs were extracted with frame length of 25 ms, frame shift of 10 ms, 512-point DFT, 23 Mel filters, pre-emphasis coefficient of 0.97, and lifter coefficient of 22. For each frame, the first 13 MFCCs were kept for a 13-dimensional feature set.

The second set of features was similar to the baseline MFCCs except that the DFT was converted to the Mel scale and normalized using f_o normalization. The default f_o was chosen to be $f_{o,def}$ = 100 Hz, and $f_{o,utt}$ was chosen to be the median f_o across the utterance computed using the MBSC pitch detection algorithm (Tan and Alwan, 2013) when f_o normalization was performed. Baseline and f_o normalized features were extracted from all utterances.

When applying f_o perturbation, $f_{o,def}$ was shifted by ±20, ±40, and ±60 Mels, resulting in $f_{o,def} \in$ {58.52, 72.10, 85.93, 100.00, 114.32, 128.90, 143.74} Hz as explained in Section 2.5. Including the initial features, this multiplies the amount of training data by 7.

Feature extraction implementation is relatively simple when using Eq. 2 to compute the effective DFT frequency bin shift. We choose $f_{o,def} \in$ {58.52, 72.10, 85.93, 100.00, 114.32, 128.90, 143.74} Hz where every value is used when performing f_o perturbation and only $f_{o,def}$ = 100 Hz is used when performing a standard feature extraction. Also, we choose $f_{o,utt}$ to be the median f_o across the utterance when performing f_o normalization and simply let $f_{o,utt}$ = 100 Hz when extracting non-normalized features. The remaining feature extraction procedure remains unchanged.

3.3. Single Word Experiments

These experiments examined the effectiveness of f_o normalization in a scenario where child speech is limited to train a single word ASR system. From the OGI single words dataset, 1,654 word utterances were randomly chosen from each grade and separated by grade for a total of 11 subsets containing approximately 1.2 hours of speech. Each subset from 1st to 10th grade was used to train several GMM-based and DNN-based single word ASR systems, either with or without normalization. For these experiments, the baseline MFCCs used a bandwidth of 8 kHz, and the f_o normalized MFCCs used a bandwidth of 6.2 kHz as explained in Section 2.6. To compare f_o normalization to other standard normalization techniques, additional child ASR systems were trained using piecewise VTLN (Lee and Rose, 1998), F3 normalization (Cui and Alwan, 2006), and normalization using the third SGR (SG3) (Guo et al., 2015). The input to the GMM-based systems was empirically chosen to be a 7-frame concatenation (3 frames left, 3 frames right) linear discriminant analysis (LDA) mapped to a 40-dimensional vector. An additional 9-frame LDA was applied to the input of the GMM-based systems for the input of the DNN-based systems.

Due to the limited amount of data, the GMMs only used a maximum of 1250 Gaussians. Similarly, the DNNs only had 2 hidden layers using 2-norm non-linearities with input dimension of 500 and output dimension of 100. The output of both the GMMs and DNNs was 250 senone probabilities.

The kindergarten subset of 1,654 words was used for testing. The language model (LM) used for decoding was a single word LM that contained the 208 possible words from the OGI single words dataset. The ASR systems trained on baseline features were tested using baseline features. Similarly, the ASR systems trained on normalized features were testing using the same normalized features.

3.4. Continuous Speech Experiments

These experiments examined the effectiveness of f_o normalization and f_o perturbation in a scenario where adult data can be used to train a continuous speech child ASR system. Adult ASR systems were first trained using the LibriSpeech data, either with or without f_o normalization. Similar to the single word experiments, the baseline MFCCs in these experiments used a bandwidth of 8 kHz, and the f_o normalized MFCCs used a bandwidth of 6.2 kHz. Senone alignments were first extracted by training GMM-HMM ASR systems using the LibriSpeech tri6b recipe from the Kaldi Speech Recognition Toolkit (Povey et al., 2011). The alignments were then used to train BLSTM-HMM ASR systems using the PyKaldi2 toolbox (Lu et al., 2019). The input to the BLSTM was empirically chosen to be a 7 frame concatenation (3 frames left, 3 frames right) for a 91-dimensional feature input. The BLSTM had 3 layers with 512 cells in each direction, followed by a feed-forward softmax layer that mapped the BLSTM output to approximately 5,700 senone probabilities.

Each adult ASR system was adapted to child speech using either the OGI sentences dataset or the CMU Kids Speech Corpus. All the parameters of the adult acoustic model were used as an initialization for training the child acoustic model, and the same training procedure was applied using child speech data for parameter fine-tuning. Approximately 70% of the data from these child speech datasets was used as adaptation data. The ASR systems trained on baseline MFCCs were adapted using baseline MFCCs. Similarly, the ASR systems trained on f_o normalized MFCCs were adapted using f_o normalized MFCCs. Additionally, data augmentation using f_o perturbation was also applied to the child adaptation data.

To compare f_o perturbation to a standard data augmentation technique, we applied vocal tract length perturbation (VTLP) (Jaitly and Hinton, 2013) to the baseline MFCCs extracted from the child speech adaptation data. This was used to adapt the adult ASR systems trained on baseline MFCCs. For a fair comparison, the set of warping factors for VTLP was chosen to be {0.94, 0.96, 0.98, 1.00, 1.02, 1.04, 1.06}, which multiplies the training data by 7.

The remaining 30% of utterances in the OGI sentences dataset or the CMU Kids Speech Corpus was used for testing. A 4-gram LM trained on approximately 14,500 Project Gutenburg books was used for decoding. This LM is one of the LMs included in Kaldi's LibriSpeech recipe (Panayotov et al., 2015).

4. Results and Discussion

4.1. Single Word Experiments

The results of the single word experiments for the GMM-based systems are displayed in Table 3, separated by grade. The first row shows the word error rates (WERs) of the child ASR systems trained with no normalization. The next three rows show the WERs of the systems trained with VTLN, F3 normalization, and SG3 normalization. The bottom row shows the WERs of the systems trained with f_o normalization. All WERs that are not significantly different ($p > 0.05$) from the f_o-based normalization are in bold.

As the results were tested on kindergarten speech, the youngest group from the OGI Kids' Speech Corpus, we expect that training data from younger children would be better matched to the testing data and produce better performance. Indeed, as the training grade increases, the performance dramatically decreases. Unsurprisingly, both f_o normalization and VTLN did not affect performance when training on 1st to 6th grade children. That is, since the children from these younger grades are physiologically similar to kindergarten children, their speech acoustic spaces are likely well-matched even without warping. As such, normalization will have a limited effect.

When training using speech from 7th grade children and older, the normalization begins to have a more pronounced effect on the system. Starting from 8th grade, systems using f_o normalization performed significantly better ($p < 0.05$) than systems using no normalization with a relative improvement of 11.8%, 13.3%, and 22.3% for 8th grade, 9th grade, and 10th grade, respectively. This is expected due to the speech acoustic mismatch between young and older children. Furthermore, this implies that the f_o normalization procedure was successful in mapping older and younger child speech to a more similar acoustic space.

The results of the single word experiments for the DNN-based systems are shown in Table 4. Similar to the GMM-based systems, both f_o normalization and VTLN did not affect performance when training on speech from 1st to 6th grade children. Starting from 7th grade, systems using f_o normalization performed significantly better ($p < 0.05$) than systems using no normalization with a relative improvement of 11.1%, 19.3%, 20.9%, and 21.0% for 7th grade, 8th grade, 9th grade, and 10th grade, respectively.

The f_o normalization technique outperformed VTLN significantly ($p < 0.05$) when the GMM-based system was trained on 10th grade child speech and when the DNN-based system was trained on 8th grade child speech. This suggests that f_o normalization is slightly more effective than VTLN at normalizing speech to some default space while using less bandwidth. Moreover, f_o normalization has an additional computational advantage as VTLN requires multiple passes through the system while f_o normalization only requires one pass. As such, the f_o normalization method is capable of providing additional computational efficiency with no loss in ASR performance.

Similarly, the f_o normalization technique outperformed F3 and SG3 normalization significantly for almost all the training grades, regardless of whether a GMM-based system or DNN-based system was used. In fact, F3 and SG3 normalization generally performed worse than the baseline unless the systems were trained on older child speech. Previous studies on these normalization techniques did not show such performance degradation (Cui and Alwan, 2006; Guo et al., 2015; Yeung and Alwan, 2019). This may be attributed to the fact that those studies trained ASR systems using a bandwidth of 4 kHz. As average F3 and SG3 values are generally between 2 kHz and 4 kHz, applying F3 or SG3 normalization with a 4 kHz bandwidth signal normalizes a majority of the frequency content. However, applying these normalization techniques to an 8 kHz bandwidth signal may distort the frequency content above the average F3 or SG3 values causing performance degradation. Regardless, this suggests that f_o normalization is a more suitable feature normalization technique for bandwidths greater than 4 kHz than F3 or SG3 normalization.

A few implications arise from the results of these experiments. Mainly, regardless of the acoustic model used, f_o normalization is the most effective at improving ASR performance on young child speech out of the feature normalization techniques tested, particularly when training data consists of older children or adults. This would similarly hold true if the training data contains a mixture of older and younger speakers. As f_o normalization is far less computationally expensive than VTLN, this further justifies the importance of an effective single-pass normalization technique. However, f_o normalization alone is not able to fully compensate for extreme acoustic mismatches, as evidenced by the results in Tables 3 and 4.

4.2. Continuous Speech Experiments

The results of the BLSTM-based system trained using f_o perturbation for data augmentation, along with the system using VTLP, are shown in Table 5. Applying f_o perturbation to the ASR system using OGI sentences results in a substantial improvement from 6.84% to 5.85%, and this result is significant ($p < 0.001$). However, the ASR system using CMU Kids shows less improvement. When applying VTLP instead of f_o perturbation, the OGI sentences system also results in an improvement but still performs worse than f_o perturbation. Additionally, the CMU Kids system performs worse than the baseline when applying VTLP. This result suggests that the f_o perturbation method is superior to VTLP since f_o perturbation creates additional data that preserves the acoustic properties of speech. This further implies that when performing frequency warping for data augmentation, the warping functions must be constructed in a way that produces realistic variations in speech features.

The results of the continuous speech experiments using both f_o normalization and f_o perturbation are shown in Table 6. The left two columns indicate whether f_o normalization or f_o perturbation were applied to the ASR system. The right two columns display the WERs of the ASR systems trained and tested on CMU Kids or OGI sentences.

When applying f_o normalization to the ASR system, we observe a slight improvement for the OGI sentences system, but

Table 3: Word error rates (WERs) (%) of GMM-HMM ASR systems for the single word experiments. Each ASR system was trained on a single grade level (1st-10th grade) and tested on kindergarten speech from the OGI single words dataset. MFCCs were extracted with no normalization, VTLN, F3 normalization, SG3 normalization, and f_o normalization. All WERs that are not significantly different ($p > 0.05$) from the f_o-based normalization are in **bold**.

Feature Normalization	Training Grade 1	2	3	4	5	6	7	8	9	10
None	**23.28**	**22.91**	**22.85**	**25.51**	**25.70**	**28.72**	**31.86**	36.64	39.66	44.74
VTLN	**24.55**	**22.31**	**23.64**	**24.83**	**23.34**	**27.51**	**29.14**	**31.86**	**34.10**	38.21
F3	27.03	26.42	28.66	**27.57**	27.75	31.92	**31.32**	**33.01**	**35.25**	38.75
SG3	**25.76**	25.15	**25.76**	**26.06**	**25.39**	**28.72**	**31.26**	**31.92**	**35.97**	**36.46**
f_o Norm	**24.55**	**22.85**	**24.61**	**26.42**	**25.21**	**27.87**	**29.81**	**32.29**	**34.40**	**34.76**

Table 4: Word error rates (WERs) (%) of DNN-HMM ASR systems for the single word experiments. Each ASR system was trained on a single grade level (1st-10th grade) and tested on kindergarten speech from the OGI single words dataset. MFCCs were extracted with no normalization, VTLN, F3 normalization, SG3 normalization, and f_o normalization. All WERs that are not significantly different ($p > 0.05$) from the f_o-based normalization are in **bold**.

Feature Normalization	Training Grade 1	2	3	4	5	6	7	8	9	10
None	**20.07**	**20.56**	**21.22**	**23.82**	**22.19**	**26.12**	28.76	32.22	36.22	41.23
VTLN	**19.35**	**19.59**	**20.62**	**22.37**	**20.68**	**23.82**	**26.18**	29.69	**29.56**	**34.64**
F3	24.55	24.12	24.30	**25.27**	25.15	29.99	28.90	29.69	**30.89**	35.85
SG3	24.24	23.10	24.61	**24.85**	24.18	**27.21**	30.29	29.14	32.71	**35.13**
f_o Norm	**20.07**	**20.25**	**20.74**	**24.61**	**21.22**	**24.43**	**25.57**	**26.00**	**28.66**	**32.56**

this improvement is not as substantial as applying f_o perturbation without f_o normalization. Applying f_o normalization along with f_o perturbation provides further improvement to 5.52%, resulting in a relative improvement over the baseline of 19.3%. When applying both f_o normalization and f_o perturbation on the CMU Kids system, the WER decreases to 16.47%, a relative improvement of 2.4%. Both systems result in improvements over recently reported child ASR systems such as Wu et al. (2019), which reported WERs of 10.8% when using OGI Kids' scripted speech and 17.3% when using CMU Kids.

The discrepancy between the improvements of the two systems is likely due to the age range of the child speakers. The OGI sentences dataset used children from kindergarten to 5th grade. Meanwhile, the CMU Kids corpus only used children from 1st to 3rd grade, excluding two outlier speakers. As seen in Section 4.1, f_o normalization is only effective when the age range of the training and testing data is large enough. This may also hold true of f_o perturbation. That is, since there is less variability in the CMU Kids corpus, adding additional variability in the training data through f_o perturbation is both unnecessary and unhelpful when training the BLSTM. Meanwhile, since the OGI sentences dataset has a larger age variability, both f_o normalization and f_o perturbation provided improvements.

To further analyze the effectiveness of the f_o normalization technique on a BLSTM-based ASR system, we applied f_o normalization to the clean and noisy adult speech testing datasets included in the LibriSpeech corpus ("test_clean", "test_other") and evaluated the effectiveness of f_o normalization using the unadapted BLSTMs, that is, systems trained only on the LibriSpeech data. The system with no normalization achieved 5.47% and 14.83% WER on the clean and noisy test sets, respectively. The system with f_o normalization achieved 5.60% and 14.87% WER on the clean and noisy test sets, respectively. These results seem to suggest that f_o normalization has virtually no effect on adult ASR. However, this implies that f_o normalization can be used in an ASR system to improve the performance on child speech with no sacrifice to the performance on adult speech.

To further analyze the results of Table 6, we separated the WERs of the OGI sentences system by grade with and without f_o normalization and f_o perturbation, shown in Table 7. The child ASR systems in Table 7 are the same as in Table 6. Similar to the previous results, the system with both f_o normalization and f_o perturbation performs better than the baseline system as well as the systems using only f_o normalization or only f_o perturbation for all grades present in the experiment. This system was significantly better than the baseline for kindergarten, 1st grade, and 2nd grade speech at $p < 0.05$. Meanwhile, the system was significantly better than the baseline for 3rd grade and 4th grade speech at $p < 0.1$. As such, we can conclude that both f_o normalization and f_o perturbation provide improvements for the entire age range for the OGI Kids' Speech Corpus.

5. Conclusion

This paper makes two contributions to the fields of child speech acoustics and child ASR. First, it reports an examination of the tonotopic distances between formants and f_o and their applications to ASR. The tonotopic distances were reformulated as a linear relationship between formants and f_o in the perceptual frequency scale. To verify this relationship, the first three formants and f_o were measured from utterances of children say-

Table 5: Word error rates (WERs) (%) of the child ASR experiment using a BLSTM-based acoustic model adapted from adult speech. ASR systems either used no data augmentation, VTLP, or f_o perturbation. WERs for both CMU Kids and OGI sentences are reported. The system using f_o perturbation performed the best on both datasets. The best performing system that performed significantly better than the baseline ($p < 0.05$) is in **bold**.

Augmentation	CMU Kids	OGI Sent.
None	16.88	6.84
VTLP	17.05	6.22
f_o Per.	16.63	**5.85**

Table 6: Word error rates (WERs) (%) of the child ASR experiment using a BLSTM-based acoustic model adapted from adult speech. The left two columns indicate whether f_o normalization ("f_o Norm?") and data augmentation using f_o perturbation ("f_o Per?") were used. WERs for both CMU Kids and OGI sentences are reported in the latter columns. The system using both f_o normalization and f_o perturbation performed the best on both datasets. The best performing system that performed significantly better than the baseline ($p < 0.05$) is in **bold**.

f_o Norm?	f_o Per?	CMU Kids	OGI Sent.
No	No	16.88	6.84
Yes	No	16.93	6.50
No	Yes	16.63	5.85
Yes	Yes	16.47	**5.52**

ing hVd words, and least squares linear regression lines relating the formant frequencies to f_o were computed. The regressions were significant with $p < 0.001$ and $r > 0.51$, consistent with the reformulation of the tonotopic distances.

This result prompted the second contribution, which was a spectral warping technique inspired by the relationship between formants and f_o in the vowel space as defined by these tonotopic distances. By choosing to warp the entire spectrum rather than just the formants, the warping technique relies on only the median f_o across the utterance, $f_{o,utt}$, and a target speech acoustic space to serve as a default speaker, defined by $f_{o,def}$. Frequency normalization is performed simply by choosing a fixed $f_{o,def}$ while data augmentation can be applied by choosing multiple values of $f_{o,def}$ in a reasonable way.

The frequency normalization and data augmentation procedures, denoted as f_o normalization and f_o perturbation, respectively, were verified on both a limited data single word child ASR experiment and a continuous speech ASR experiment. The single word ASR experiment explored the effectiveness of f_o normalization and compared the results against other normalization techniques. When the child ASR system was trained on older child speech and tested on kindergarten speech, f_o normalization performend significantly better than the baseline and slightly better than the system using VTLN. The continuous speech ASR experiment explored the effectivenes of f_o normalization and f_o perturbation and found that the combination of both techniques was beneficial to a BLSTM-based child ASR system when the age range of the child speakers was large. Additionally, f_o perturbation was found to be more helpful than VTLP as a data augmentation technique. Finally, an examination of the ASR performance on individual grade levels revealed that the combination of f_o normalization and f_o perturbation was beneficial across a range of ages. The results of this paper further suggest the importance of frequency warping using a warping function that adheres to the physical constraints of speech. Furthermore, the experiments performed demonstrate the effectiveness of the f_o warping techniques regardless of the acoustic model used.

While the proposed f_o normalization and f_o perturbation techniques provide significant improvements to child ASR systems, there is a number of additional research directions to explore. Further improvement of child ASR is still necessary, especially for the youngest children. Another direction of particular importance is the development of a universal ASR system for both adults and children.

Acknowledgements

This work was funded in part by National Science Foundation (NSF) Grant #1734380.

References

Assmann, P.F., Nearey, T.M., 2007. Relationship Between Fundamental and Formant Frequencies in Voice Preference. The Journal of the Acoustical Society of America 122, EL35–EL43. doi:10.1121/1.2719045.

Barreda, S., Nearey, T.M., 2012. The Direct and Indirect Roles of Fundamental Frequency in Vowel Perception. J. Acoust. Soc. Am. 131, 466–477. doi:10.1121/1.3662068.

Barreda, S., Nearey, T.M., 2013. The Perception of Formant-Frequency Range is Affected by Veridical and Judged Fundamental Frequency, in: Proc. Meetings on Acoustics, p. 060197. doi:10.1121/1.4800915.

Bunnell, H.T., Yarrington, D.M., Polikoff, J.B., 2000. STAR: Articulation Training for Young Children, in: Proc. ICSLP, pp. 85–88.

Chistovich, L.A., 1985. Central auditory processing of peripheral vowel spectra. The Journal of the Acoustical Society of America 77, 789–805. doi:10.1121/1.392049.

Chistovich, L.A., Lublinskaya, V.V., 1979. The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. Hearing Research 1, 185–195.

Cui, X., Alwan, A., 2005. MLLR-Like Speaker Adaptation Based on Linearization of VTLN with MFCC Features, in: Proc. INTERSPEECH, pp. 273–276.

Cui, X., Alwan, A., 2006. Adaptation of Children's Speech with Limited Data Based on Formant-Like Peak Alignment. Comput. Speech Lang. 20, 400–419. doi:10.1016/j.csl.2005.05.004.

Cui, X., Goel, V., Kingsbury, B., 2014. Data augmentation for deep neural network acoustic modeling, in: Proc. ICASSP, pp. 5582–5586. doi:10.1109/ICASSP.2014.6854671.

Eskenazi, M., Mostow, J., Graff, D., 1997. The CMU Kids Speech Corpus LDC97S63. URL: https://catalog.ldc.upenn.edu/LDC97S63.

Fahey, R.P., Diehl, R.L., Traunmüller, H., 1996. Perception of Back Vowels: Effects of Varying F1-F0 Bark Distance. J. Acoust. Soc. Am. 99, 2350–2357.

Fainberg, J., Bell, P., Lincoln, M., Renals, S., 2016. Improving children's speech recognition through out-of-domain data augmentation, in: Proc. of INTERSPEECH, pp. 1598–1602.

Faria, A., Gelbart, D., 2005. Efficient Pitch-Based Estimation of VTLN Warp Factors, in: Proc. of INTERSPEECH, pp. 213–216.

Fujinaga, K., Nakai, M., Shimodaira, H., Sagayama, S., 2001. Multiple-Regression Hidden Markov Model, in: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 513–516.

Gerosa, M., Giuliani, D., Narayanan, S., Potamianos, A., 2009. A Review of ASR Technologies for Children's Speech, in: Proc. WOCCI, pp. 7.1–7.8. doi:10.1145/1640377.1640384.

Table 7: Word error rates (WERs) (%) of the child ASR experiment using a BLSTM-based acoustic model adapted from adult speech. The left two columns indicate whether f_o normalization ("f_o Norm?") and data augmentation using f_o perturbation ("f_o Per?") were used. WERs on OGI sentences, separated by testing grade, are reported in the latter columns. The system using both f_o normalization and f_o perturbation performed the best for all grades. The best performing system that performed significantly better than the baseline ($p < 0.05$) is in **bold**.

		Testing Grade					
f_o **Norm?**	f_o **Per?**	K	1	2	3	4	5
No	No	16.97	9.17	6.73	5.71	4.15	4.99
Yes	No	17.44	9.17	6.19	4.80	3.47	4.93
No	Yes	13.97	7.89	5.27	5.11	3.72	4.35
Yes	Yes	**12.87**	**7.38**	**4.88**	4.78	3.35	4.24

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S., 2014. A pitch extraction algorithm tuned for automatic speech recognition, in: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2494–2498.

Gray, S.S., Willett, D., Lu, J., Pinto, J., Maergner, P., Bodenstab, N., 2014. Child Automatic Speech Recognition for US English: Child Interaction with Living-Room-Electronic-Devices, in: Proc. WOCCI, pp. 21–26.

Guo, J., Paturi, R., Yeung, G., Lulich, S.M., Arsikere, H., Alwan, A., 2015. Age-Dependent Height Estimation and Speaker Normalization for Children's Speech Using the First Three Subglottal Resonances, in: Proc. INTERSPEECH, pp. 1665–1669.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., Ng, A.Y., 2014. Deep Speech: Scaling up end-to-end speech recognition. arXiv:arXiv:1412.5567v2.

Jaitly, N., Hinton, G.E., 2013. Vocal tract length perturbation (VTLP) improves speech recognition, in: Proc. ICML.

Kathania, H.K., Kadiri, S.R., Alku, P., Kurimo, M., 2020. Study of Formant Modification for Children ASR, in: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE. pp. 7429–7433.

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., Belpaeme, T., 2017. Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations, in: Proc. ACM/IEEE HRI, pp. 82–90.

Kewley-Port, D., Watson, C.S., Elbert, M., Maki, D., Reed, D., 1991. The Indiana Speech Training Aid (ISTRA) II: Training Curriculum and Selected Case Studies. Clin. Linguist. Phon. 5, 13–38. doi:10.3109/02699209108985500.

Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., Khudanpur, S., 2017. A study on data augmentation of reverberant speech for robust speech recognition, in: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 5220–5224.

Koenig, L.L., Lucero, J.C., 2008. Stop Consonant Voicing and Intraoral Pressure Contours in Women and Children. J. Acoust. Soc. Am. 123, 1077–1088. doi:10.1109/TMI.2012.2196707.Separate, arXiv:NIHMS150003.

Koenig, L.L., Lucero, J.C., Perlman, E., 2008. Speech Production Variability in Fricatives of Children and Adults: Results of Functional Data Analysis. J. Acoust. Soc. Am. 124, 3158–3170. doi:10.1121/1.2981639.

Lee, L., Rose, R., 1998. A Frequency Warping Approach to Speaker Normalization. IEEE Transactions on Speech and Audio Processing 6, 49–60.

Lee, S., Potamianos, A., Narayanan, S., 1997. Analysis of Children's Speech: Duration, Pitch and Formants, in: Proc. EUROSPEECH, pp. 473–476.

Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of Children's Speech: Developmental Changes of Temporal and Spectral Parameters. J. Acoust. Soc. Am. 105, 1455–1468.

Ljolje, A., 2002. Speech Recognition Using Fundamental Frequency and Voicing in Acoustic Modeling, in: Proc. of the International Conference on Spoken Language Processing (ICSLP), pp. 2137–2140.

Lu, L., Xiao, X., Chen, Z., Gong, Y., 2019. PyKaldi2: Yet another speech toolkit based on Kaldi and PyTorch. arXiv:1907.05955 arXiv:arXiv:1907.05955.

Magimai-Doss, M., Stephenson, T.A., Bourlard, H., 2003. Using Pitch Frequency Information in Speech Recognition, in: Proc. of EUROSPEECH, pp. 2525–2528.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: An ASR corpus based on public domain audio books, in: Proc. IEEE ICASSP, pp. 5206–5210. doi:10.1109/ICASSP.2015.7178964.

Panchapagesan, S., Alwan, A., 2006. Multi-Parameter Frequency Warping for VTLN by Gradient Search, in: Proc. IEEE ICASSP, pp. 1181–1184.

Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019. SpecAugment: A simple data augmentation method for automatic speech recognition, in: Proc. INTERSPEECH, pp. 2613–2617. doi:10.21437/Interspeech.2019-2680.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., Veselý, K., 2011. The Kaldi Speech Recognition Toolkit, in: Proc. IEEE ASRU.

Sadeghian, R., Zahorian, S.A., 2015. Towards an Automated Screening Tool for Pediatric Speech Delay, in: Proc. INTERSPEECH, pp. 1650–1654.

Serizel, R., Giuliani, D., 2014. Vocal Tract Length Normalization Approaches to DNN-Based Children's and Adults' Speech Recognition, in: Proc. SLT, pp. 135–140.

Shahnawazuddin, S., Dey, A., Sinha, R., 2016. Pitch-Adaptive Front-End Features for Robust Children's ASR, in: Proc. of INTERSPEECH, pp. 3459–3463. doi:10.21437/Interspeech.2016-1020.

Sheng, P., Yang, Z., Qian, Y., 2019. GANs for Children: A Generative Data Augmentation Strategy for Children Speech Recognition, in: Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 129–135.

Shivakumar, P.G., Georgiou, P., 2020. Transfer Learning From Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations. Computer Speech and Language 63. doi:10.1016/j.csl.2020.101077.

Shivakumar, P.G., Potamianos, A., Lee, S., Narayanan, S., 2014. Improving Speech Recognition for Children Using Acoustic Adaptation and Pronunciation Modeling, in: Proc. WOCCI, pp. 15–19.

Shobaki, K., Hosom, J.P., Cole, R.A., 2000. The OGI Kids' Speech Corpus and Recognizers, in: Proc. ICSLP, pp. 258–261.

Smith, B.L., 1992. Relationships Between Duration and Temporal Variability in Children's Speech. J. Acoust. Soc. Am. 91, 2165–2174. doi:10.1121/1.403675.

Spaulding, S., Chen, H., Ali, S., Kulinski, M., Breazeal, C., 2018. A Social Robot System for Modeling Children's Word Pronunciation, in: Proc. AAMAS, pp. 1658–1666.

Stemmer, G., Hacker, C., Steidl, S., Nöth, E., 2003. Acoustic Normalization of Children's Speech, in: Proc. EUROSPEECH, pp. 1313–1316.

Syrdal, A.K., Gopal, H.S., 1986. A Perceptual Model of Vowel Recognition Based on the Auditory Representation of American English Vowels. J. Acoust. Soc. Am. 79, 1086–1100. doi:10.1121/1.393381.

Tan, L.N., Alwan, A., 2013. Multi-Band Summary Correlogram-Based Pitch Detection for Noisy Speech. Speech Comm. 55, 841–856. doi:10.1016/j.specom.2013.03.001.

Tepperman, J., Silva, J., Kazemzadeh, A., You, H., Lee, S., Alwan, A., Narayanan, S., 2006. Pronunciation Verification of Children's Speech for Automatic Literacy Assessment, in: Proc. INTERSPEECH, pp. 845–848.

Traunmüller, H., 1981. Perceptual Dimension of Openness in Vowels. J. Acoust. Soc. Am. 69, 1465–1475. doi:10.1121/1.385780.

Vorperian, H.K., Kent, R.D., 2007. Vowel Acoustic Space Development in Children: A Synthesis of Acoustic and Anatomic Data. J. Speech. Lang. Hear. 50, 1510–1545. doi:10.1044/1092-4388(2007/104).Vowel.

Wu, F., 2020. Child Speech Recognition As Low-Resource Automatic Speech Recognition. M.S. thesis, The Johns Hopkins University. URL: https://jscholarship.library.jhu.edu/handle/1774.2/62766.

Wu, F., Garcia, L.P., Povey, D., Khudanpur, S., 2019. Advances in automatic speech recognition for child speech using factored time delay neural network, in: Proc. INTERSPEECH, pp. 1–5. doi:10.21437/interspeech.2019-2980.

Yeung, G., Afshan, A., Ozgun, K.E., Kaewtip, K., Lulich, S.M., Alwan, A., 2017. Predicting Clinical Evaluations of Children's Speech with Limited Data Using Exemplar Word Template References, in: Proc. SLaTE, pp. 161–166.

Yeung, G., Alwan, A., 2018. On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children, in: Proc. INTERSPEECH, pp. 1661–1665.

Yeung, G., Alwan, A., 2019. A Frequency Normalization Technique for Kindergarten Speech Recognition Inspired by the Role of fo in Vowel Perception, in: Proc. INTERSPEECH, pp. 6–10. doi:10.21437/interspeech.2019-1847.

Yeung, G., Bailey, A.L., Afshan, A., Pérez, M.Q., Martin, A., Spaulding, S., Park, H.W., Alwan, A., Breazeal, C., 2019. Towards the Development of Personalized Learning Companion Robots for Early Speech and Language Assessment, in: Proc. AERA.

Yeung, G., Lulich, S.M., Guo, J., Sommers, M.S., Alwan, A., 2018. Subglottal Resonances of American English Speaking Children. The Journal of the Acoustical Society of America 144, 3437–3449. doi:10.1121/1.5082289.

- Fundamental frequency-based feature warping formulated from relationship between formants and f_o
- Warping technique can be used for normalization or data augmentation
- Applying both f_o-based normalization and data augmentation to child ASR provides a substantial improvement over baseline

- Gary Yeung – First Author
- Ruchao Fan – Second Author
- Abeer Alwan – Third Author

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: