Speaking-Rate Adaptation of Automatic Speech Recognition System through Fuzzy Classification based Time-Scale Modification

S. Shahnawazuddin
Dept. of ECE
NIT Patna
s.syed@nitp.ac.in

Hemant K. Kathania
Dept. of ECE NIT Sikkim
India
hemant.ece@nitsikkim.ac.in

Nagaraj Adiga
Dept. of CS University of Crete
Greece
nagaraj@csd.uoc.gr

B. Tarun Sai
Dept. of ECE NIT Patna
India
s.syed@nitp.ac.in

Waquar Ahmad
Dept. of ECE NIT Calicut
India
waquar@nitc.ac.in

***Abstract*—In this paper, we study the role of speaking-rate adaptation (SRA) of automatic speech recognition (ASR) systems. The performance of an ASR system is reported to degrade when the speaking-rate is either too fast or too slow. In order to simulate such a situation, an ASR system was trained on adults' speech and used for transcribing speech data from adult as well as child speakers. Earlier studies have shown that, speaking-rate is significantly lower in the case of children when compared to adults. Consequently, the recognition performance for children's speech was noted to be very poor in contrast to adults' speech. To improve the recognition performance with respect to children's speech, speaking-rate was explicitly changed using time-scale modification (TSM). A recently proposed TSM approach based on fuzzy classification of spectral bins has been explored in this regard. The fuzzy-classification-based TSM technique is reported to be superior to state-of-the-art approaches. Effectiveness of the said TSM technique has not been studied yet in the context of ASR. The experimental studies presented in this paper show that SRA based on fuzzy classification results in a relative improvement of 30% over the baseline.**

***Index Terms*—Speaking-rate adaptation, automatic speech recognition, time-scale modification, fuzzy classification.**

I. Introduction

The task of modifying the duration of a signal without changing the frequency contents is referred to as time-scale modification (TSM). Several different techniques for TSM have been proposed over the years [1]. Changing the duration of a signal is beneficial for numerous applications. Common examples are modifying the length of a music signal in order to fit within a prescribed time-slot and viewing a video in slow-motion. Another important area of signal processing where TSM finds application is the task of changing the speaking-rate (SR). Different people speak at different rates, some being slow while others being fast. Consequently, it is difficult to recognize the spoken words when the speaking-rate is extremely fast. The same problem is faced by automatic speech recognition (ASR) systems as well. Even though an ASR system is expected to be insensitive towards variations in speaking-rate, earlier studies have shown that the recognition performance degrades significantly when the speaking-rate is too fast or too slow [2]–[5]. Hence, TSM can be used to enhance the robustness of ASR systems towards speaking-rate variations.

In this paper, we have experimentally studied the ill-effects of speaking-rate variation on the recognition performance of an ASR system. To simulate such a task, an ASR system was developed using speech data from adult speakers. Next, the developed system was used for transcribing speech data from adult as well as child speakers. The task of decoding children's speech using acoustic models trained on adults' data is an example where the differences in the speaking-rate are highly pronounced. In the case of children, the average phoneme duration is longer [4], [6], [7]. Hence, the speaking-rate for children is lower than that for adults. In addition to that, variability in speaking-rate is higher among the child speakers themselves. Consequently, the error rates were noted to be much higher in the case of children's speech when compared to task of decoding adults' data.

In order to normalize the differences in speaking-rate, we have studied the role of TSM in this work. In other words, to improve the recognition performance with respect to children's speech, TSM has been employed for suitably increasing the speaking-rate for child speakers. The approach for TSM explored in this study is the one based on fuzzy classification of spectral bins [8]. The fuzzy-classification-based TSM technique has been proposed recently and is reported to be better than the existing similar approaches. Its effectiveness in the context of ASR system has not been studied yet. Speaking-rate adaptation (SRA) via fuzzy-classification-based TSM is noted to be highly effective as demonstrated by the experimental studies presented in this paper. Even though TSM was explored in some of the earlier reported works [9], [10], acoustic modeling based on Gaussian mixture models (GMM) was employed in those works. In this paper, we have used acoustic

978-1-5386-9286-8/19/$31.00

modeling approaches based deep neural networks (DNN) [11] and long short-term-memory (LSTM) [12] recurrent neural networks (RNN) for experimental evaluations.

The rest of this paper is organized as follows: In Section II, the proposed speaking-rate adaptation technique is described. In Section III, the experimental evaluations demonstrating the effectiveness of the proposed approach are presented. Finally, the paper is concluded in Section IV.

II. Speaking-rate adaptation

A. Motivation

As stated earlier, speaking-rate for adult and child speakers differ significantly. It was observed in [13]–[15] that, production as well as perception of phones get affected by the variation in speaking-rate. Therefore, when speaking-rate is exceptionally fast or slow, the recognition performance of an ASR system is observed to degrade severely. A typical example of such a scenario is the task of decoding speech data from children using an ASR system trained on adults' speech. The average vowel durations in the case of children's speech are longer than those for the adults. An increase in average vowel duration implies that the speaking-rate for children is lesser than that for adults [4].

In order to gain a better understanding about the aforementioned differences, speaking-rate was computed for adults and children using a large number of continuous speech utterances from both the groups of speakers. The speech data used for this purpose was derived from two separate **British English** corpora, namely WSJCAM0 [16] (adults' speech) and PF-STAR [17] (children's speech). Further details about those speech databases are summarized later in Section III. In this work, we have quantified the speaking-rate in terms of number of phonemes per second. The variation of speaking-rate in the case of adult and child speakers is demonstrated using the histograms shown in Figure 1. We have used 500 utterances from each group of speakers in order to derived those histograms. It is evident from Figure 1 that, the mean speaking-rate for adult speakers is almost two times greater than that for the children. Since data driven machine learning techniques are commonly used for the task of speech recognition, extreme acoustic mismatch occurs when an ASR system trained on adults' speech is used for transcribing children's data. In order to improve the recognition performance, speaking-rate normalization through TSM was explored in some of the earlier reported works on children's ASR [9], [10]. Pitch-synchronous overlap and add (PSOLA) algorithm was used for SRA in [9]. On the other hand, pitch-synchronous time-scaling [18] was studied in [10]. Motivated by those works, we have also explored SRA for improving children's speech recognition in this paper. In this regard, a recently proposed TSM approach based on fuzzy classification of spectral bins is explored. The fuzzy-classification-based technique for TSM is reported to be superior to state-of-the-art approaches.

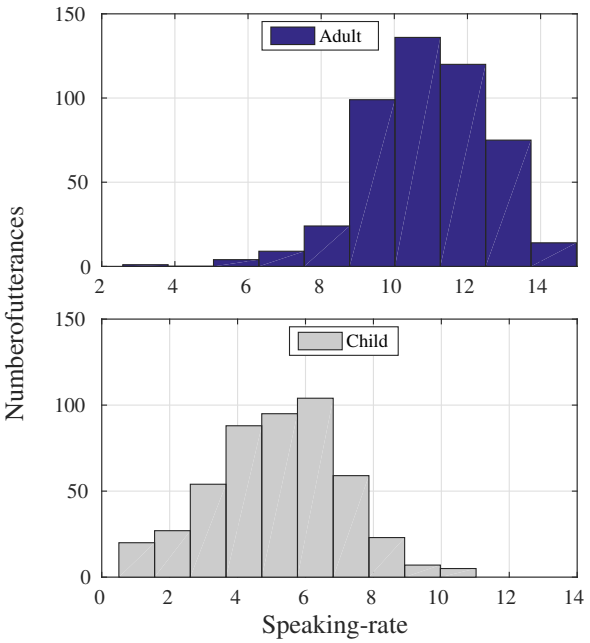

Fig. 1: Histograms depicting the variation in speaking-rate for adult and child speakers quantified in terms of number of phonemes per second. Speaking-rate was computed using 500 utterances from adult and children.

B. Fuzzy classification based time-scale modification

Any audio signal can be considered to consist of three different components, viz. sinusoidal, noise, and transient [19]. In the context of TSM, the most challenging part is to preserve the quality of those three components simultaneously. To deal with this challenge, some of the earlier reported TSM approaches resorted to a binary classification of the spectral bins. However, binary classification has a serious drawback since the energy in each spectral bin is actually a combination of energy from each of those three aforementioned components [8]. To overcome this limitation, the spectral bins should belong to each of the three classes at the same time with an associated degree of class-membership. In other words, the approach for classifying the spectral bins should be *fuzzy* [20] instead of being binary. Motivated by this fact, the characteristics of an audio signal were quantified using fuzzy classification in a recently proposed technique for TSM [8]. When compared to the approach based on harmonic-percussive separation [21], fuzzy-classification-based TSM was observed to be better. Motivated by its success, we have explored the effectiveness of this technique in the context of automatic speech recognition.

C. Effect of TSM on speaking-rate

The effect of TSM on the duration of given speech signal is shown through a set of time domain waveforms in Figure 2. In order to increase the speaking, the given speech signal was compressed by a factor of 0.7. On the other, a scaling factor of 1.4 was used for decreasing the speaking-rate. The corresponding spectrograms are also shown in Figure 2. From the figure, it is evident that the shape of the speech signal as well as the formant transitions are preserved even after

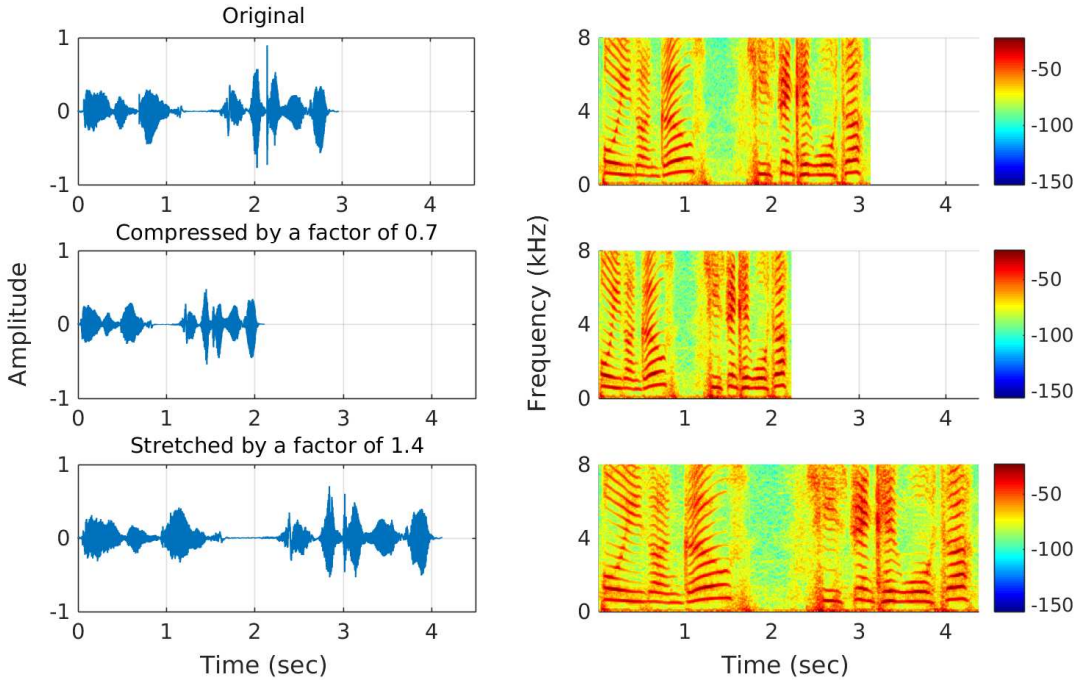

Fig. 2: The time-domain waveforms for a segment of speech and the corresponding compressed and stretched versions. The spectrograms for each of the cases are also shown. It is note that, the shape of the signal as well as formant transitions remain preserved even after time-scale modification.

increasing/decreasing the speaking-rate. Next, the speaking-rate of the children's speech used for the analysis presented in Figure 1 was increased by modifying the duration of each of the utterances using a factor of 0.7. The histogram depicting the variation of speaking-rate after modification are shown in Figure 3 (bottom pane). For proper contrast, the histogram obtained prior to TSM is also shown Figure 3 (top pane). By comparing the two histograms, an increase in mean speaking-rate is noticeable. It can be concluded from these analyses that, by optimally compressing the signal duration, the acoustic mismatch resulting from the differences in speaking-rate can be reduced to a large extent. In the next section, the simulation studies performed to validate this claim are presented in detail.

III. Experimental results and discussions

In this section, we first describe the experimental setup employed in this work. This is followed by the experimental evaluations validating the effectiveness of speaking-rate adaptation.

A. Experimental setup

Speech Corpora: Adults' speech data used in this work was obtained from WSJCAM0 [16] British English speech corpus for continuous speech recognition. A train set (Adult-Train) was derived from WSJCAM0 for learning the statistical model parameters. Adult-Train set consisted of 15.5 hours of speech data from 92 adult male and female speakers. Further, the train set comprised of 132,778 words and the total number of utterances was 7852. A test set (Adult-Test) was also derived in order to measure the matched case recognition

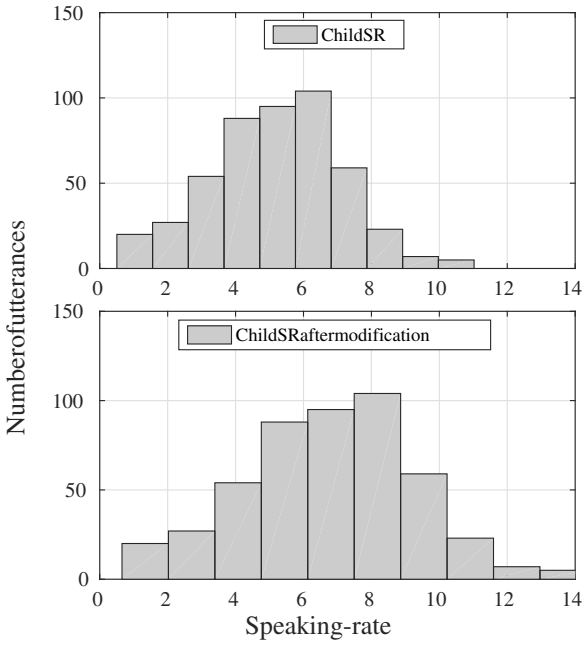

Fig. 3: Histogram depicting variation in speaking-rate for children before and after time-scale modification. Speaking-rate is quantified in terms of number of phonemes per second

performance. The Adult-Test set consisted of 0.6 hours of speech data from 20 speakers with a total of 5,608 words. In order to simulate the mismatched ASR task discussed earlier, children's speech data was obtained from PF-STAR corpus [17]. Like WSJCAM0, PF-STAR corpus is also a British English speech database. A test set (Child-Test) was

derived from PF-STAR corpus which consisted of speech data from 60 children aged between 3 to 14 years. The total duration of speech data was 1.1 hours. There were a total of 5067 words in Child-Test data set. The experimental studies reported in this work were performed on a wide-band speech data sampled at a rate of 16 kHz. *Front-end speech parameterization:* The speech data was first pre- emphasized using a high-pass filter. The pre-emphasis factor was selected as 0.97. Next, frame-blocking was done using overlapping Hamming windows of length 20 ms with an overlap of 50%. In other words, the frame-overlap was chosen to be 10 ms. In order to extract the 13-dimensional base Mel-frequency cepstral coefficients (MFCC), a 40-channel Mel-filter bank was employed. The 13-dimensional base MFCC features were then spliced in time taking a context size of 9 frames. Time-splicing resulted in 117-dimensional vectors which were then reduced to 40-dimensional features vectors using linear discriminant analysis (LDA) and maximum-likelihood linear transformation (MLLT). Cepstral mean and variance normalization (CMVN) as well as feature-space maximum-likelihood linear regression (fMLLR) were performed next to enhance the robustness towards speaker-dependent variations. The required fMLLR transformations for the training and test data were generated through speaker adaptive training. *Specifications of the ASR system:* The ASR systems were developed on the 15.5 hours adults' speech data from the WSJCAM0 speech corpus using the Kaldi toolkit [22]. Context-dependent hidden Markov models (HMM) were used for modeling the cross-word triphones. Decision tree-based state tying was performed with the maximum number of tied-states (senones) being fixed at 2000. Acoustic modeling based on deep neural network and the long short-term-memory recurrent neural network (RNN) was explored as stated earlier. Prior to learning parameters of the DNN-HMM-based ASR system, the fMLLR-normalized feature vectors were time-spliced once again considering a context size of 9 frames. The number of hidden layers in the DNN was chosen as 8 with each layer consisting of 1024 hidden nodes. The nonlinearity in the hidden layers was modeled using the *tanh* function. The initial learning rate for training the DNN-HMM parameters was set at 0.005 which was reduced to 0.0005 in 15 epochs. The minibatch size for neural net training was selected as 512. The LSTM-based acoustic models were trained with 4 hidden layers each having 1024 nodes. The dimension of the LSTM cell was chosen as 1024. The number of epochs used for LSTM training was set to 5 while the initial and final learning rates were selected to be 0.005 and 0.0005, respectively.

While evaluating the matched case performance or decoding the Adult-Test set, the MIT-Lincoln 5k Wall Street Journal bi-gram language model (LM) was used. The perplexity of this LM for the Adult-Test set is 95.3 while there are no out-of-vocabulary (OOV) words. Further, a lexicon consisting of 5,850 words including pronunciation variants was used. While decoding the Child-Test set, a 1.5k domain-specific bigram LM was used. This bigram LM was trained on the transcripts of speech data in PF-STAR after excluding those corresponding to Child-Test set. The said domain-specific LM has an OOV rate of 1.20% and perplexity of 95.8 for the Child-Test set. The lexicon used while decoding the Child-Test set consisted of 1,969 words including pronunciation variations.

TABLE I: Baseline WERs for Child-Test and Adult-Test sets on adult data trained DNN- and LSTM-based ASR systems.

Acoustic modeling technique	WER (in %)		Relative difference (%)
	Child-Test	Adult-Test	
DNN	19.27	5.87	70
LSTM	16.33	5.10	69

B. Baseline recognition performances

The word error rate (WER) metric was employed to measure the recognition performance. The baseline WERs for Child-Test and Adult-Test datasets with respect to the adult data trained DNN- and LSTM-based ASR systems are given in Table I. The ill-effects of aforementioned factors of acoustic mismatch can be easily understood by noting the large difference in WERs for the two sets. It is to note that, the presented WERs were obtained after applying CMVN and fMLLR in order to reduce the speaker-dependent acoustic mismatch. In addition to that, domain-specific LMs were used while decoding the corresponding test sets. Yet, the recognition performance for children's speech is much poorer compared to adults' case. The relative difference in WERs for the two test sets highlight this point.

C. Effect of speaking-rate adaptation

In order to change the speaking-rate for children's speech, the TSM factor was varied from 0.65 to 1.35 in steps of 0.05. Modification factor values less than unity imply an increase in speaking-rate. On the other hand, in order to decrease the speaking-rate, values greater than one are used. The correspondingly modified test data was then decoded to improve the recognition rates. The WER profiles demonstrating the effect increasing and decreasing the speaking-rate are shown in Figure 4. Since the speaking-rate is lower in the case children's speech, the WER is observed to decrease when the scaling factor is less than unity. Similar trends are noted for both DNN- as well as LSTM-based ASR systems. The best case WERs along with the percentage relative improvement over the baseline obtained through SRA are given in Table II. From these results it is evident that proposed approach for SRA is extremely effective.

IV. Conclusion

The role of speaking-rate adaptation in the context of automatic speech recognition has been studied in this work. In this regard, a recently proposed time-scale modification technique based on fuzzy classification of spectral bins is explored. This fuzzy-classification-based TSM approach is reported to be better than the existing similar techniques. In order to simulate

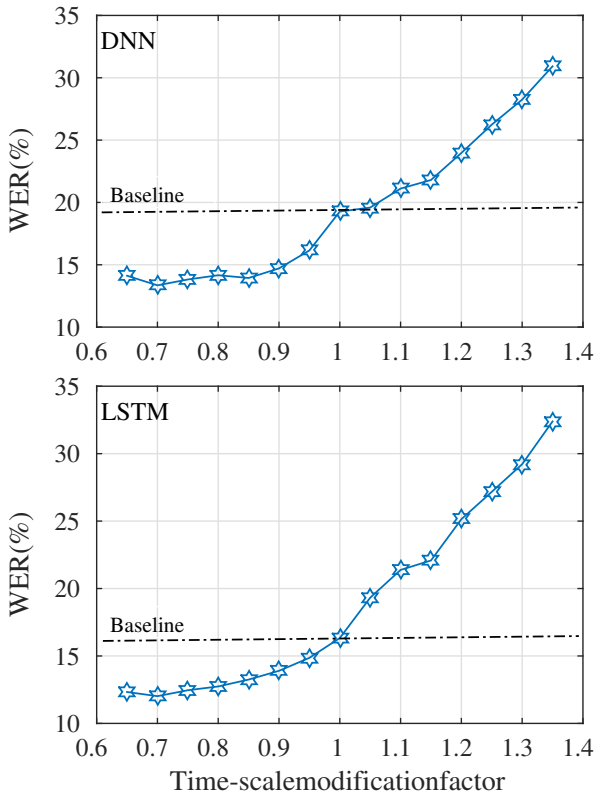

Fig. 4: WERs illustrating the effect of increasing/decreasing the speaking-rate on the recognition of children's speech using DNN- and LSTM-based ASR system trained on adults' speech.

TABLE II: The best case WERs for children's speech test set obtained through SRA.

Acoustic	WER (in %)		Relative
model	Baseline	SRA	improvement (%)
DNN	19.27	13.35	30.7
LSTM	16.33	12.02	26.4

an ASR task where large differences in speaking-rate exists, children's speech is transcribed using acoustic models trained on speech data from adult speakers. Significant reductions in word-error rates are obtained by suitably changing the speaking-rate through fuzzy-classification-based TSM.

REFERENCES

[1] J. Driedger and M. Müller, "A review of time-scale modification of music signals," *Applied Sciences*, vol. 6, no. 2, p. 57, 2016.

[2] R. Kent and L. Forner, "Speech Segment Durations in Sentence Recitations by Children and Adults," *Journal of Phonetics*, vol. 8, pp. 157–168, 1980.

[3] S. Lee, A. Potamianos, and S. S. Narayanan, "Analysis of children's speech: Duration, pitch and formants," in *Proc. INTERSPEECH*, vol. 1, September 1997, pp. 473–476.

[4] ——, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, March 1999.

[5] S. M. Chu and D. Povey, "Speaking rate adaptation using continuous frame rate normalization," in *Proc. ICASSP*, March 2010, pp. 4306–4309.

[6] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Proc. Speech and Language Technologies in Education (SLaTE)*, September 2007.

[7] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proc. Workshop on Child, Computer and Interaction*, 2009, pp. 7:1–7:8.

[8] E.-P. Damskägg and V. Välimäki, "Audio time stretching using fuzzy classification of spectral bins," *Applied Sciences*, vol. 7, no. 12, p. 1293, 2017.

[9] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth, "Acoustic normalization of childrens speech," in *Proc. INTERSPEECH*, September 2003, pp. 1313–1316.

[10] S. Ghai, "Addressing Pitch Mismatch for Children's Automatic Speech Recognition," Ph.D. dissertation, Department of EEE, Indian Institute of Technology Guwahati, India, October 2011.

[11] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] J. L. Miller, "Effects of speaking rate on segmental distinctions," *Perspectives on the study of speech*, pp. 39–71, 1981.

[14] Q. Summerfield, "Articulatory rate and perceptual constancy in phonetic perception," *Journal of Experimental Psychology: Human Performance and Perception*, vol. 7, pp. 208–215, 1981.

[15] J. L. Miller and L. E. Volaitis, "Effect of speaking rate on the perceptual structure of a phonetic category," *Perception & Psychophysics*, vol. 46, no. 6, pp. 505–512, November 1989.

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

[17] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, and M. Wong, "The PF_STAR children's speech corpus," in *Proc. INTERSPEECH*, 2005, pp. 2761–2764.

[18] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for prosodic and voice quality transformations," in *Proc. INTERSPEECH*, 2005, pp. 1137–1140.

[19] T. S. Verma and T. H. Meng, "An analysis/synthesis tool for transient signals that allows a flexible sines+ transients+ noise model for audio," in *Proc. ICASSP*, vol. 6, 1998, pp. 3573–3576.

[20] L. A. Zadeh, "Making computers think like people," *IEEE spectrum*, vol. 21, no. 8, pp. 26–32, 1984.

[21] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 105–109, 2014.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, December 2011.