Rapid Unsupervised Adaptation to Children's Speech on a Connected-Digit Task

Daniel C. Burnett and Mark Fanty

Center for Spoken Language Understanding
Oregon Graduate Institute of Science & Technology

ABSTRACT

We are exploring ways in which to rapidly adapt our neural network classifiers to new speakers and conditions using very small amounts of speech, say, one or a few words. Our approach is to perform a speaker-dependent warping of the frequency scale by selecting a *Bark offset* for each speaker. We choose the offset for a speaker to be the one that maximizes our recognizer output score on the adaptation utterance. We then use the speaker's offset during evaluation of all other utterances by the speaker. To test our approach, we evaluate an adult-speech trained recognizer on children's speech from the same task both before and after adaptation to each child's voice. Using only a single digit for adaptation, we have reduced the word error rate for children's speech from 9.6% to 4.2%. Using a seven-digit utterance further reduced the error rate to 3.5%.

1. Introduction

It is important that real-world speech recognition systems work for as many people as possible; otherwise someone may be denied a service or a business may lose a customer. However, even high-performance recognition systems occasionally encounter individual speakers for whom performance is low. We feel that an important strategy for achieving the maximum user throughput is to rapidly (during one session) adapt the recognizer. Such rapid adaptation will necessarily be based on a small amount of data—perhaps a single word.

A common adaptation approach for HMM-based systems is to adjust the means and variances of the various models, but in principle this requires several examples for each model, an impossibility when the adaptation data is a single word. Recently researchers have primarily focused on two solutions to this problem: correlation between model mean and variance differences for different speakers and more global optimization methods.

The internals of neural network based systems are not as amenable to manipulation as the means and variances of HMM systems. Speaker adaptation approaches used with neural network recognizers are varied and include (1) remapping the feature space before input to the net or remapping the net outputs, and (2) developing a mapping between an input speaker's spectra and those of a "generic" speaker which can then be used to adjust all incoming speech by that speaker.

Our goal is to develop an adaptation method that is simple, rapid, and effective for our neural network based recognizer. Specifically, we are searching for a small number of parameters which can simply and rapidly be adjusted globally based on a small amount of "local" data. In this paper we present our adaptation method, we describe a parameter that explicitly accounts for variability among speakers, and we demonstrate the effectiveness of the adaptation method and parameter on the task of rapidly adapting an adult-trained recognizer to better handle children's speech.

2. Adaptation procedure

We want our adaptation to be simple and rapid. Specifically, we intend to use a single utterance by a speaker to estimate $p_{optimal}$, the optimal value of our parameter. This parameter value will then be used to recognize all of the remaining utterances by the speaker. To keep things simple, we perform a linear minimization on the output score (a log likelihood) from our recognizer for the adaptation utterance as we vary the parameter value. We use Brent's algorithm [8, section 10.2], a cross between the golden section search and parabolic interpolation, to perform the minimization. Given starting locations a, x_0, and b such that $a \leq x_0 \leq b$, $f(x_0) \leq a$, and $f(x_0) \leq b$, Brent's algorithm is guaranteed to find, to a specified precision of ϵ, the location x_{min} and value y_{min} of a local minimum of a function $y = f(x)$, where $a < x_{min} < b$.

Note that the success of the above procedure depends on several assumptions:

1. Brent's algorithm requires that we specify a range [a,b] for the parameter p within which our recognizer score $R(u,p)$ has a single minimum.
2. For speed purposes, we must be able to estimate $p_{optimal}$ with a small number of recognizer passes over the utterance.

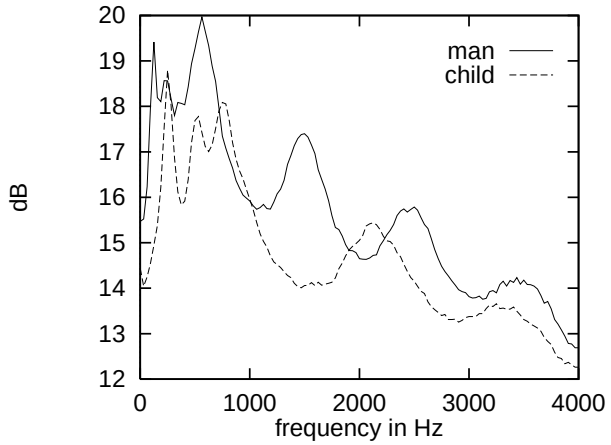

Figure 1: Averaged power spectra for the vowel /ɛ/ from males and children, presented on the Hz scale.

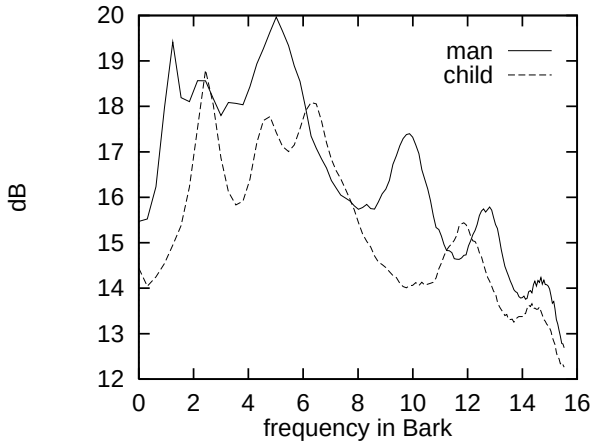

Figure 2: Averaged power spectra for the vowel /ɛ/ from males and children, presented on the Bark scale.

3. Our estimate of $p_{optimal}$ must be accurate enough to improve the recognizer performance on the remaining utterances for the speaker. Note that this implies that the score produced by our recognizer is fairly well correlated with the accuracy of the recognizer.

3. Bark-scale offset parameter

One factor that varies among speakers, especially between adults and children, is the length of the individual's vocal tract. One effect of the difference in vocal tract lengths can be seen in the speech spectra of vowels. Bladon, *et al.* [2], compared average spectra, for the vowel /ɛ/, for male and female speakers of Northern British English and noticed that when presented on the Bark scale (a 'tonality' scale, based on the frequency analysis properties of the ear's basilar membrane, which expands the resolution of lower frequencies and compresses that of higher frequencies, the spectra differed, on average, only in their location on the Bark scale. They hypothesized that by displacing all frequency components of the average female vowel downward by one unit on the Bark scale, one would obtain the average spectrum of the same male vowel. A similar comparison between male speakers and children is shown in figures 1 through 3.

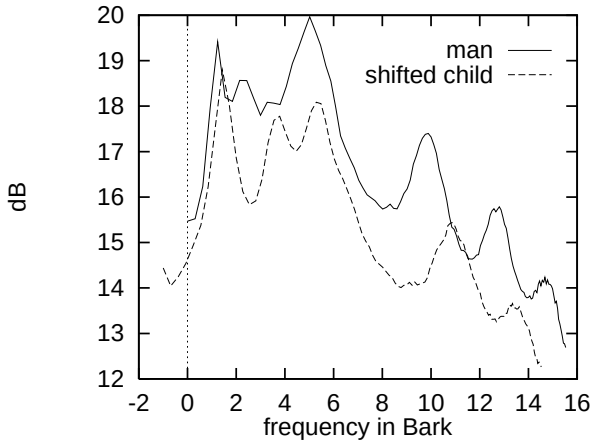

Figure 3: Averaged power spectra for the vowel /ɛ/ from males and children, presented on the Bark scale. The children's spectra have been uniformly shifted left by 1.5 Bark.

The idea of warping the frequency scale to normalize for different vocal tracts is not new. Several recent papers have reported on experiments with various scalings of the frequency axis. Both Eide and Gish [3] and Lee and Rose [6] have proposed extensions to the work of Andreou, *et al.* [1]. This work has also been extended by the authors [5], who experimented with linearly scaling the frequency axis for speakers from the Switchboard corpus. The optimal scale factor was selected by trying different values and choosing the one that maximized the a-posteriori probability of the data. Although most of the improvement occurred in this first step, they also iteratively retrained their HMM by selecting an optimal scale factor for each training speaker, retraining using those scale factors, reselecting the optimal scale factor for each speaker, etc. Their performance improved from a baseline of 49% correct (46% accuracy) to 54% correct (51% accuracy). Similarly, Eide and Gish experimented with a simple exponential warping (designed to allow more adjustment at high frequencies than at low frequencies) of the speakers' frequency scales. However, to avoid the computational load of evaluating the system at many (arbitrary) warping parameter values for each speaker, they attempted to estimate the parameter value based on third formant values for the speaker relative to those for other speakers. Their error rates, also on Switchboard, showed an 8-10% drop for several different dataset sizes and conditions. Lee and Rose propose HMM-based procedures for estimating an appropriate scaling factor and describe a simple implementation of the frequency warping using a direct modification of the filters in their front-end. Experiments on a telephone-based connected digit recognition task demonstrated a noticeable

reduction in error rate. Another approach based on an explicit voiced speech model is given by Wegmann, *et al.* [9]. Their piecewise-linear mapping produced a 12% reduction in error rates on a Switchboard task.

For our experiments, we have decided to approximate this difference between speakers' spectra by a linear shift (offset) of all frequency components on the Bark scale, under the assumption that there is a single ideal offset for each speaker that is appropriate for all utterances by that speaker.

4. Speech Recognition System

Our speech recognition system can be broken down into the following four stages: signal processing (spectral analysis), feature selection, neural network classification, and word search.

Signal Processing. The 8kHz digitized speech is temporally sliced into frames of 10ms width, each of which is run through a seventh order PLP analysis to obtain 8 cepstral coefficients. This step is described in more detail in section 4.1.

Feature Selection. For each frame, the 8 coefficients from that frame, along with the 8 coefficients from each of six nearby frames (covering a total of approx. 160 ms of speech), are connected to form a 56-element feature vector for the frame.

Neural Network Classifier. Each vector is classified by a 3 layer feed-forward artificial neural network with 200 hidden units into one of 209 categories, each of which corresponds to a phoneme in a given left and right context.

Word Search. The frame outputs from the ANN are used, along with duration and word pronunciation constraints, in a Viterbi search to produce the recognized string. The language model allows the string to consist of one or more of the digits plus "oh," optionally separated by silence.

4.1. Bark offset parameter

Our PLP analysis follows that proposed by Hermansky[4]: Hamming-windowed FFT, warping of power spectrum to the Bark scale, critical-band masking, equal-loudness preemphasis, intensity-loudness conversion, then LPC cepstral analysis.

In order to allow for the Bark offset parameter, we have rewritten the Hz-to-Bark transformation to be

$$\Omega(f) \quad = \quad 6\ln\left\{\frac{f}{600}+\sqrt{\left(\frac{f}{600}\right)^2+1}\right\}+\textbf{offset}$$

Note that this merely redefines which spectral values (in Hz) will be summed together in each of the critical-band smoothing functions.

5. Adaptation Task

Since the recognizer was trained exclusively on adult speech, we test on children's speech to test the ability of our technique to adjust for vocal tract length differences.

5.1. Datasets

Our baseline classifier was trained using adult utterances from the "train" subset of the TIDIGITS corpus[7], a corpus of studio-quality connected-digit utterances. In this corpus, each speaker (adult or child) spoke 22 single-digit utterances and 11 each of 2-, 3-, 4-, 5-, and 7-digit utterances, for a total of 77 utterances per speaker. Because our recognizers are designed for telephone speech, we downsampled the TIDIGITS data from the original 20kHz sampling rate to the desired 8kHz rate. Since our system requires time-aligned phonetic labels for training, we bootstrapped by using an existing number recognizer to label the training data, trained the new classifier, relabeled using the new classifier, and so on until there was no significant improvement in the baseline performance. The baseline performance for adult speakers was 99.3% word accuracy.

Adaptation and testing were performed on the children's utterances from the "train" subset of the TIDIGITS corpus. There were a total of 51 children in this subset.

5.2. Experiments

A preliminary analysis of the data indicated that essentially all of the optimal Bark offset values would fall within the range [-2.0, 0.0]. In addition, plots of the variation in recognizer score within this range as a function of Bark offset showed a fairly smoothly-varying function with a clear minimum region. These two indications satisfied the first of the 3 assumptions given in section 2.

As a baseline for our experiments, we tested each of the 51 children in the training set on that child's utterances using an offset of 0.0, then averaged together these digit-level accuracy scores to obtain the baseline error rate shown in Table 1.

In our first experiment, we performed the following procedure for each child: For each 7-digit utterance (out of 11 spoken by the child), we used Brent's algorithm to obtain the offset that maximized the recognizer score for that utterance, then used that offset to test, for digit-level accuracy (using the NIST total-sub-ins-del scoring method), the remaining 76 utterances spoken by the child. We then found the average of these eleven accuracy scores. We then computed the average, across all of the children, of these averaged accuracy scores, obtaining the overall system error rate.

Our second experiment followed exactly the same process as the first, but in it we adapted using the children's single-digit utterances. Results from both experiments are presented in Table 1.

Condition		*Word Error Rate*
Baseline		9.6%
Adapted	***1-digit string***	4.2%
	7-digit string	3.5%

Table 1: Baseline and after-adaptation results when adapting using either a single-digit string or a seven-digit string.

6. Discussion

In this paper, we have presented a promising approach for adapting to children's speech that is

Simple We have identified a simple, well-motivated single parameter that is easy to add to our system, and our method of selecting the ideal parameter value is just a simple, common linear minimization method.

Rapid In practice, only 8-10 passes of the recognizer over a single 1-digit utterance are necessary to find a good estimate of the optimal offset for a child.

Effective For a system trained only on adult speakers, this simple adaptation reduced the error rate on the children's speech by more than 50%.

There are several points we should consider. Some of these we have begun to analyze, while others still remain open for further exploration: There is a clear gap between the adult and child (adapted) error rates. Two obvious suspects are the accuracy of our estimate of $p_{optimal}$ and the optimality of our warping function. However, it is likely that there are prosodic or other differences between the speech of adults and children.

The Bark shift was motivated by the differences in formant locations within voiced regions of speech. However, unvoiced speech (fricatives, silence, etc.) is in general less affected by vocal tract length. The ideal shift, then, would only be computed on voiced regions of speech and likewise only be applied to voiced regions. For simplicity we have ignored this complication, but it should be examined.

This adaptation was *unsupervised* in the sense that we did not know what digits the child spoke in the adaptation utterance, although we did know that it was a string of digits. We feel that this is therefore a more powerful result than had we done supervised adaptation, but it remains to be seen whether or not supervised adaptation would produce better performance.

Our adaptation approach estimates an optimal offset for each speaker using a recognizer trained on speech from speakers for whom we did not find the optimal offset. This mismatch between the training and evaluation conditions can possibly be corrected through an iterative training procedure similar to that done by Kamm, *et al.* [5]. In our case we would use our existing recognizer to find the optimal offset for each training speaker, regenerate the neural network's training examples, retrain, redetermine the optimal offset for each training speaker, and so on until we obtain no further improvement. We would expect that the performance on children's speech would then be worse without adaptation but better with it, since the system now *expects* to be given appropriately normalized data. Note that the adaptation process itself would take the same amount of time – the only difference is that now adaptation would be required instead of optional.

7. Acknowledgments

This work was supported by grants from the National Science Foundation, the Defense Advanced Research Projects Agency, and the member companies of CSLU.

8. REFERENCES

1. Andreas Andreou, Terri Kamm, and Jordan Cohen. Experiments in vocal tract normalization. In *Proceedings of the CAIP Workshop: Frontiers in Speech Recognition II*, 1994.
2. R. A. W. Bladon, C. G. Henton, and J. B. Pickering. Towards an auditory theory of speaker normalization. *Language & Communication*, 4(1):59–69, 1984.
3. Ellen Eide and Herbert Gish. A parametric approach to vocal tract length normalization. In *Proceedings of ICASSP 96*, volume 1, pages 346–349, 1996.
4. Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.
5. Terri Kamm, Andreas G. Andreou, and Jordan Cohen. Vocal tract normalization in speech recognition: Compensating for systematic speaker variability. In *Proceedings XV Speech Research Symposium*, Johns Hopkins University, 1995.
6. Li Lee and Richard C. Rose. Speaker normalization using efficient frequency warping procedures. In *Proceedings of ICASSP 96*, volume 1, pages 353–356, 1996.
7. R. G. Leonard. A database for speaker-independent digit recognition. In *Proceedings of ICASSP 84*, volume 3, page 42.11, 1984.
8. William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, second edition, 1994.
9. Steven Wegmann, Don McAllaster, Jeremy Orloff, and Barbara Peskin. Speaker normalization on conversational telephone speech. In *Proceedings of ICASSP 96*, volume 1, pages 339–341, 1996.