# VOICE TRANSFORMATIONS FOR IMPROVING CHILDREN'S SPEECH RECOGNITION IN A PUBLICLY AVAILABLE DIALOGUE SYSTEM

*Joakim Gustafson[1] and Kåre Sjölander[2]*

[1]Telia Research AB
Farsta, Sweden
joakim.k.gustafson@telia.se

[2]Centre for Speech Technology (CTT)
KTH, Stockholm,Sweden
kare@speech.kth.se

## ABSTRACT

To be able to build acoustic models for children, that can be used in spoken dialogue systems, speech data has to be collected. Commercial recognizers available for Swedish are trained on adult speech, which makes them less suitable for children's computer-directed speech. This paper describes some experiments with on-the-fly voice transformation of children's speech. Two transformation methods were tested, one inspired by the Phase Vocoder algorithm and another by the Time-Domain Pitch-Synchronous Overlap-Add (TD-PSOLA) algorithm. The speech signal is transformed before being sent to the speech recognizer for adult speech. Our results show that this method reduces the error rates in the order of thirty to forty-five percent for children users.

## 1. INTRODUCTION

Commercially available speech recognizers of today are mostly trained on adult speech, making them less suitable for spoken input from children. Thus, when designing children-directed spoken dialogue applications, it is important to collect data under realistic circumstances in order to construct language models that work well for children. However, simple data collection methods that are used to collect adult speech, like reading prepared sentences from a paper, or having people talk to a simple system-directed dialogue system, do not work that well for children. It is problematic to record children reading aloud [1], and having children talk to a dialogue system fails because of the very problem mentioned in the beginning - the recognition performance is bad for children, making it hard to get a dialogue going. This is a bootstrapping problem - how do you collect spoken dialogue data from children (in order to improve recognition for children) without having a good recognizer available? In previous studies this problem has been solved using Wizard-of-Oz techniques to collect dialogue data from children [2,6]. However, if the dialogue system is to be used in a permanent public exhibition this solution is too expensive and complicated.

This article describes a method that aims at solving this problem for a dialogue system placed in a permanent exhibition. The idea is to transform the children's speech to lower frequencies before down-sampling it to telephone bandwidth. This method has been implemented in a simple system directed dialogue system that was designed to be engaging for children. It is a small speech enabled computer game publicly available in a futuristic apartment that is on display at the Telecommunication museum in Stockholm. The users interact with an animated agent to fix things in the apartment or to get to know more about the agent personally.

The results when applying our method on the speech data collected with this system show that we can reduce the error rates between thirty and forty-five percent.

## 2. RELATED RESEARCH

Previous studies have investigated the problems with speech recognition for children and have indicated that more training data from children may not be the only solution. In a digit recognition experiment on Danish telephone speech the error rates for children were 170% higher than for adults [3]. By increasing the amount of training data for children this difference was reduced, but still 100% worse. One reason for this might be that there is more information above the telephone bandwidth in child speech than in adult speech [4]. In this study, HTK was used to train acoustic wideband models for equal size training sets from a database of read speech by children and the TIMIT database with adult speakers. The baseline error rates were 21% for children and 35% for the adults. They then decreased the bandwidth of the wideband training sets in steps of 2kHz and retrained the recognizers. The error rates increased relatively little for both groups until about 4kHz where the error rate grew more rapidly for the children. Their conclusion was that telephone bandwidth is worse for recognition on children's speech than on speech from adults.

Speaker normalization techniques have proved successful in improving speech recognition rates for both adults and children [5-8]. Vocal Tract Normalization (VTN) schemes typically try to compensate for differences in vocal tract length between different speakers, which is a major source of inter-speaker variability. The differences in formant frequencies between male adults and children can be up to 50%. The normalization is done through warping of the spectral envelope. This can be implemented efficiently during the feature extraction stage by varying spacing and width of individual filters of the mel-scale filter bank. VTN can be applied both during training and during recognition. During training speaker dependent frequency warping functions are applied to create "normalized" acoustic models. For recognition an utterance dependent warping function is chosen, which subsequently is used when computing feature vectors. Obviously this necessitates detailed control over both training and decoding procedures.

A previous study has shown that the use of frequency warping to normalize children's speech before training of acoustical models can reduce the recognition error rates by up to 55% [7]. Frequency warping was also used in another study that showed substantial improvement in children's performance on a command-and-control speech recognition system trained using adult speech [8].


Figure 1: *The animated talking agent Pixie, an overview of the exhibition and some of the young users interacting with Pixie.*

## 3. THE PIXIE SYSTEM

To be able to build conversational systems for all ages it is important to collect realistic spoken dialogue data in public environments. The August system was our first attempt to do this. It was a spoken dialogue system where users could interact with the animated agent August [9]. Since it was a walk-up use-once system it was both hard to determine what the users' goals were and to gather user-related information, such as sex and age. In the current Pixie system these problems were addressed by making all users register before interacting with Pixie, and by having both system-directed dialogues and dialogues where the users are given the initiative.

Telia Research has been involved in setting up an exhibition called "Tänk Om" ("What If") at the Telecommunication museum in Stockholm. It is a permanent exhibition with a full-size apartment of the year 2010. It tries to visualize some technical concepts that might be found in home environments in the future. Among other things the visitors can interact with the animated agent Pixie, shown to the left in Figure 1. Pixie is supposed to visualize an embodied speech interface to both information services and home control. The visitors are asked to either help Pixie to perform certain tasks in the apartment or to ask Pixie general questions about her self or the exhibition. The last will be referred to as the social dialogue scenario.

### 3.1. The logistics of the exhibition

The exhibition is presented in form of a show every hour during the afternoons. The visitors can book themselves to shows that take up to 30 persons. When they arrive they are directed to the log-in area, marked in the middle picture in Figure 1. They get a smart card and a code that they use to identify themselves when logging in at any of the terminals available. They are asked to supply information about their age, length and sex as well as answer a couple of questions about their beliefs and experiences of information technology.

Next the visitors enter the cinema located in the middle of the exhibition area. They are shown an eight-minute movie in form of a sitcom about a family that lives in 2010. The movie introduces Pixie as the family's personal digital assistant. The sitcom ends in chaos and then the movie screens are raised and the visitors can enter the family's apartment that they just saw in the movie. A museum guide tells them to interact with Pixie to help her to put things straight again.

### 3.2. The spoken dialogue system

In the apartment there are twelve touch screens in the walls and tables where Pixie flies by. The users can get in contact with Pixie by inserting their smart cards, and speak in the hand-held singing microphone. This makes it possible for the system to retrieve the user information collected during login as well as a list of the tasks that the user has solved at other terminals in the apartment. The age information is used to decide when to use voice transformation.

Each of the twelve touch screens is connected to a PC-computer that runs a game server, a Nuance recognition server and a communication server that handles the information flow. The latter sends commands and information between the game server, the recognition server and the exhibition's central server. The central server can both control physical devices in the apartment and access the database that contains all information about the users and their previous interactions.

The computer game featuring Pixie was developed by the game company Liquid Media. The game consists of five small assignments where Pixie first introduces a problem and then asks the user what to do. This could either be to fix problems that were introduced in the movie or to control the home environment, e.g. change the lighting in the bedroom. Apart from these assignments Pixie can respond to about a hundred questions about herself and the exhibition.

The system uses the Nuance 7.0.4 speech recognizer with Swedish telephone bandwidth models (www.nuance.com). We built a customized audio provider that controls the input audio stream, thus making it possible to transform the sound before streaming it to the recognizer. It also makes it possible to save the speech files in wide band quality. Finally, the audio provider can provide information about the recognized utterance, such as average level or pitch. Figure 2 shows which processing steps take place in the audio provider.

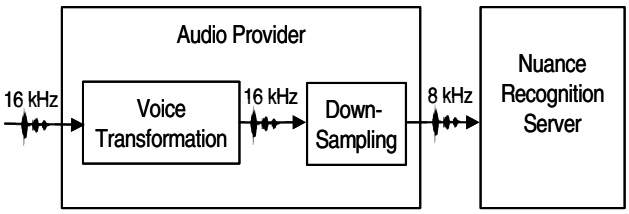


Figure 2: *The order of the speech transformations.*

## 4. VOICE TRANSFORMATION FOR CHILDREN'S SPEECH

None of the methods for Vocal Tract Normalization referred to in section 2 could be applied in the current system because of the use of a commercial 'black-box' speech recognizer. The only option was to transform the sound on-the-fly before it entered the speech recognizer.

In order to improve the system's recognition performance on children's speech several methods for voice transformation were investigated. Two promising methods were chosen for implementation and evaluation. The first, *method A*, was inspired by the Phase Vocoder algorithm [10] and the second, *method B*, was inspired by the Time-Domain Pitch-Synchronous Overlap-Add (TD-PSOLA) algorithm [11]. Method B also included linear compression of the spectral envelope of each window to give some form of formant scaling possibility.

A separate database was used for development consisting of 360 utterances spoken by ten kids in the ages 10 to 13. The sentences were constructed so that they would comply with the recognition grammar of the social dialogue scenario. The same grammar was also used with the Nuance batch recognition tool during evaluation. For *method A,* a spectral compression of 20% proved to yield the best performance. When using method B, a 23% decrease in pitch and a corresponding spectral compression gave the best results. The transformations were carried out on sound sampled at 16kHz. After transformation the signal was down-sampled to 8kHz before being fed to the speech recognizer, which used telephone-bandwidth acoustic models, see Figure 2. Since children's voices reaches higher frequencies than adults, this scheme could make it possible to utilize some of the spectral information that would be lost if using narrow-band recording.

Both methods provide transformations of reasonably high acoustic quality. However some artifacts are introduced, which result in unavoidable degradation of the signal. Method A adds a varying degree of reverberation depending on the compression factor. *Method B* is sensitive to the performance of the pitch tracker, which in this case has to work reliably for the higher-than-adult pitch range of children's voices.

The voice transformation methods were implemented and optimized using the Snack Sound Toolkit [12], which has good support for scripting, batch processing and rapid prototyping.

It is also possible to modify the rate of speech using both of the tested methods. In order to implement this in the system a robust on-the-fly method for measuring speaking rate would have to be devised. Rate transformation was not investigated in the present study, although it is known that children tend to speak more slowly than adults [13].

## 5. THE CHILDREN'S SPEECH TEST SET

The Pixie dialogue system has been publicly available seven days a week since the opening. During the period 25th of January to 12th of March almost 1,400 visitors had interacted with Pixie saying on average 11 utterances each. The users' ages have ranged between three and eighty-three. About one fourth of the users are children under the age of fifteen, but these are responsible for one third of all collected utterances.

Two test sets from the collected speech corpus were selected to investigate if the voice transformation methods would improve the recognition rates. The first set consisted of system directed dialogues where the users could change the lighting in the bedroom by selecting one of five predefined settings. The speech recognition grammars in this sub-dialogue had on average 280 states and 400 transitions. The other set consisted of user directed dialogues, where the users had been told to ask Pixie either personal questions or questions about the exhibition. The speech recognition grammar in this social scenario had about 1500 states and 2000 transitions.

Utterances that should be covered by these grammars were selected from all children up to the age of twelve years. This gave 340 lighting utterances and 335 social utterances from a total of 174 children. To be able to compare the results for the children's utterances with adults, 536 lighting utterances and 346 social utterances where selected from 327 adults users. Finally, a set of 240 utterances for the ages thirteen to fifteen was also used to investigate if the recognizer's confidence score could be used in cases where the methods made the recognition rates worse.

## 6. RESULTS

The Pixie system was used in a very difficult environment where the users interacted with Pixie simultaneously on different terminals that sometimes were as close as one meter from each other. As previously reported, word error rate might not be the best thing to measure in spoken dialogue systems displayed in acoustically challenging environments [14]. Lamel et. al. propose that it is more relevant to measure query understanding rate. Hence, in this study we decided to measure both these error rates. Table 1 shows the results of transforming the children's speech with the respective voice modification methods. As can be seen, the query error rate is reduced by about forty-five percent in the lighting case and thirty percent in the social scenario.

| Transform | Lighting | | Social | |
|---|---|---|---|---|
| | WE | QE | WE | QE |
| None | 43 | 20 | 45 | 45 |
| Method A | 31 | 11 | 36 | 32 |
| Method B | 35 | 13 | 36 | 30 |
| Adults | 19 | 5 | 28 | 24 |

*Table 1*: The word error rates (WE) and query error rates (QE) in percent for the lighting and social dialogues.

To investigate if the method worked differently for younger children the data was divided into a group of 267 utterances from three to nine-year-olds and another group of 407 utterances from ten to twelve-year-olds. The results are summarized in Table 2.

| Transform | 3-9 years old | | 10-12 years old | |
|---|---|---|---|---|
| | WE | QE | WE | QE |
| None | 59 | 39 | 36 | 28 |
| Method A | 41 | 22 | 31 | 21 |
| Method B | 44 | 24 | 31 | 21 |

*Table 2*: The word error rates (WE) and query error rates (QE) in percent for the two age groups.

As can be seen in this table, both methods are more effective for younger children, a difference that was significant according to an ANOVA analysis. The break point in our test data was at the age of twelve. For adolescents the methods makes the recognition rates higher. To use the methods you must thus know the age of the subject. A solution would be to run two recognizers in parallel, where the audio provider for one of them transforms the input speech before streaming it to its recognizer and the other not. The dialogue component would then be able to choose the output from the recognizer that got the best acoustic confidence score. This solution was tested on all 675 utterances from three to twelve-year-olds, a set of 240 utterances from thirteen to fifteen-year-olds and all 882 utterances from the adults. To see if the methods work differently for female speakers the adult set was split according to gender. About 40% of the utterances where from female users. Table 3 shows the results of applying the two methods on these test sets and how they can be adjusted by selecting the recognition result with the highest confidence score.

| Age Group | None | Method A | Method B |
|---|---|---|---|
| 3-12 years | 44 | 34 → 34 | 36 → 36 |
| 13-15 years | 37 | 44 → 36 | 38 → 37 |
| Adult female | 20 | 40 → 17 | 25 → 19 |
| Adult Male | 23 | 92 → 29 | 80 → 27 |

*Table 3*: The word error rates for the three age groups. The values to the right of the arrows are for selecting the recognition with the highest confidence score.

Both methods works equally well on speech from young children, but *method A* works worse for the older children, where it actually increases the error rates. But choosing the recognition with best confidence reduces the error rate back to the original values. *Method A* increases the error rate more than *method B*. for adults, significantly more for men than for women. Choosing the recognition with best confidence score works surprisingly good for adult women – it even decreases the error rate slightly lower than the original error rate. For men however the error rate still increases but only from 23% to 29% instead of the initial 92%.

Since previous studies concluded that children's speech has useful information above telephone bandwidth, our methods were applied before down-sampling the signal to telephone bandwidth. To verify that this actually mattered for our methods we did an experiment where they were applied after down-sampling instead of before. For *method B* this increased the error rate from 36% to 39% and for *method A* it increased from 34% to 61%. *Method A* seems to be able to benefit more from high frequency information. In any case we can conclude that it is beneficial to do the voice transformation on children's speech on 16 kHz audio and then down-sample it to telephone bandwidth.

## 7. CONCLUSIONS

A method for improving the recognition rate for children's speech in a commercial recognizer that uses telephone bandwidth acoustical models was presented. This method is used in a custom audio provider for the Nuance recognizer. In this way it is possible for the voice transformation algorithms to process wide-band speech before streaming it in telephone bandwidth to the recognizer. According to an ANOVA analysis, both transformation methods give significant reductions in error rates for children and these reductions are significantly higher for children under the age of ten. There is no significant difference in the performance on children's speech between the two methods.

A simple method of using two recognizers in parallel and choosing the one with best confidence score is also presented. However, this simple method takes processing capacity and is dependent on the recognizer's capabilities to produce usable confidence scores. A more general approach might be to develop a method that uses the acoustic signal to determine when to use the transformation method.

## 8. ACKNOWLEDGEMENTS

We would like to thank everybody who developed the "Tänk Om" exhibition, the Telecommunication museum for hosting it and their guides for handling the visitors. We also would like to thank Mattias Heldner for assistance with the variance analysis and Johan Boye and Anders Lindström for comments on drafts of this paper. Finally we would like to thank Linda Bell for her great engagement and assistance in this study.

## 9. REFERENCES

[1] Eskenazi, M., "KIDS: A database of children's speech", *Journal of the Acoustic Society of America 100*:4(2), December 1996

[2] Oviatt, S., "Talking To Thimble Jellies: Children's Conversational Speech with Animated Characters", *Proceedings of ICSLP'2000,* Vol. 3, pp. 877-880. Beijing, China, 2000.

[3] Wilpon, J. & Jacobsen, C., "A Study of Speech Recognition for Children and The Elderly", *Int. Conference on Acoustics, Speech and Signal Processing,* pp. 349-352, Atlanta, GA, 1996.

[4] Li, Q. & Russel, M., "Why is Automatic Recognition of Children's Speech Difficult?", *Proceedings of Eurospeech'01,* pp 2671-2674 Aalborg, Denmark, 2001.

[5] Lee, L. & Rose, R., "Speaker Normalization Using Efficient Frequency Warping Procedures", *Proceedings of ICASSP-96 ,* pp 353-356, 1996.

[6] Narayanan, S. & Potamianos, A., "Creating Conversational Interfaces for Children", *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2 pp. 65-77, February 2002

[7] Potamianos, A., Narayanan, S. & Lee, S., "Automatic Speech Recognition for Children", *Proceedings of Eurospeech'97,* pp 2371-2374, Rhodes, Greece, 1997.

[8] Das, S., Nix, D. & Picheny, M., "Improvements in Children's Speech Recognition Performance", *Proceedings of ICASSP' 98*, pp 433-436, Seattle, WA, 1996.

[9] Gustafson, J. & Bell, L., "Speech Technology on Trial: Experiences from the August System", *Journal of Natural Language Engineering: Special issue on Best Practice in Spoken Dialogue Systems*, 2000.

[10] Dolson, M., "The phase vocoder: A tutorial" Computer Music Journal, vol. 10, no. 4, pp. 14-27, 1986.

[11] Moulines, E. & Charpentier, F., "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", Speech Communication Vol. 9 (5/6), pp. 453-467, 1990.

[12] The Snack Sound Toolkit, http://www.speech.kth.se/snack/

[13] Lee, S., Potamianos, A. & Narayanan, S., "Analysis of Children's Speech: Duration, Pitch and Formants", *Proceedings of Eurospeech'97,* pp 473-476, Rhodes, Greece, 1997.

[14] Lamel, L., Gauvain, J., Bennacef, S., Devillers, L., Foukia, S., Gangolf, J. & Rosset, S., "Field trials of a telephone service for rail travel information", *Proceedings of IEEE Workshop on Interactive Voice Technology for Telecom. Applications*, pp.111-116, 1996.