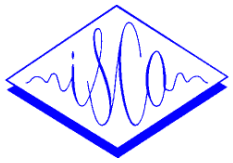
ISCA Archive
http://www.isca-speech.org/archive



ITRW on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology (PMLA2002)
Aspen Lodge, Estes Park, Colorado, USA
September 14–15, 2002

PINPOINTING PRONUNCIATION ERRORS IN CHILDREN'S SPEECH: EXAMINING THE ROLE OF THE SPEECH RECOGNIZER

Maxine Eskenazi[12], Gary Pelton[2]

[1]Language Technologies Institute, 5000 Forbes Ave. Pittsburgh, PA 15213
[2]Carnegie Speech Company, 4619 Newell Simon Hall, 5000 Forbes Ave. Pittsburgh, PA 15213
max@cs.cmu.edu, gap@cs.cmu.edu

Abstract

In speech recognition, when a system created for one application is used for another or for a different population of users, large amounts of data and engineering effort are needed to "adapt" it to its new use. Much work has recently centered on reducing that effort. This paper concerns changing from an adult to a child population of users in a system that pinpoints pronunciation errors in English. It first discusses children's speech production. Then it describes adaptation that is centered around a combination of relatively small amounts of data with minimal recognizer changes for a system that can pinpoint errors as well for children's speech as it does for adults'.
The precision of the adult system was tested on children's speech. Then Open Source SPHINX was tested on children's speech and tests were run, using a variety of parameters, that compared the precision of automatic pinpointing of recognition errors to human tutor pinpointing of errors. The various parameters tested, the test conditions, and results are discussed.

1. INTRODUCTION

In speech recognition, when a system created for one application is used for another or for a different population of users, large amounts of data and engineering effort are needed to "adapt" it to its new use. Much work has recently centered on reducing that effort. This paper concerns changing from an adult to a child population of users in a system that pinpoints pronunciation errors in English. It first discusses children's speech production. Then it describes adaptation that is centered around a combination of relatively small amounts of data with minimal recognizer changes for a system that can pinpoint errors as well for children's speech as it does for adults'. It is hoped that this information can be useful to anyone adapting a system that works for one speaker population to another, such as elderly speakers. It is also hoped that this information can shed some light on the extent to which the representation and processing of the speech signal in present automatic speech recognition systems can be used for new user populations.
We use the algorithms described briefly in [5] for the Fluency system to compare results of error detection using an adult-trained recognizer and a child-trained one. We describe the native and non-native children's data, human tutor assessment of it and subsequent tests using it.

2. A DATABASE OF CHILDREN'S SPEECH

Speech from children differs from that of adults in several ways. Very young children are still learning the sounds and expressions of their mother tongue and create very variable speech. Sounds may be under-articulated and hyperarticulated in the same utterance and pronounced in different ways each time they are produced. Elementary school children (> 6 years old) have more stable pronunciation, but their speech still differs from adult speech due to articulatory differences and vocabulary and syntax choices. The child's production system also has shorter vocal cords and vocal tract than an adult's. For all of these reasons we have chosen to collect a database of children's speech and then train the Open Source SPHINX [2] speech recognizer rather than use adult speech to recognize children's speech.

2.1 Database content

We recorded 135 children (56 boys and 79 girls) in two suburban Pittsburgh schools, These 8 – 10 year olds (good readers with no marked Pittsburgh dialect) read isolated words, phrases and sentences which were automatically chosen for their richness in triphone coverage. They are from children's texts (Mother Goose, Peter Pan, etc.).
There are an average of 188 utterances per speaker. The 179 first ones are the same for all. The 25 others were the same for every tenth child. Utterances were repeated until the recorder decided the speech corresponded to the text on the screen. If, it didn't after several attempts, then the utterance was skipped. There are 51251 seconds = 14 hr 14 min of speech, and 25,122 utterances in all.

2.2 Labelling the database

Each speech file has a text file containing the word labels of the speech. The data was not hand-labeled. On-the-fly selection described above eliminated "incorrect" utterances. After-the-fact verification was performed (listening to a random selection of the data and eliminating the utterances that did not meet this criterion - about 60 in the whole database). The low amount of rejected data shows our method of collection and implicit labeling to be successful.

3. TRAINING WITH CHILDREN'S SPEECH: TWO TYPES OF TRAINING

Two types of training, the speech recognizer (Sphinx) and the pinpointer (Fluency), are necessary to create an effective children's system. Both Sphinx and Fluency compare a child's voice to some information derived from reference voices.

3.1 Training the recognizer

We trained SPHINX on the children's data using the default settings. This gave us two recognizers (the adult system we already had and the child one) with exactly the same configurations.

The child recognizer has less triphone coverage (the adult one was trained on over 100 hours of speech – more speech from only slightly more speakers). The adult phone file has 125,715 triphones while the kids' file has only 33,195. The experience at Carnegie Mellon in building recognition systems from small amounts of data indicates that even with this small amount of data we should be able to create an effective system. In order to get better coverage, we tested the effect of changing the number of tied states. We trained four different children's systems – 6000-tied states, 4000, 3000, and 2000. Below are comparative results for the first three systems. Results for the 2000-tied state system were close to chance and are not shown.

3.2 Training the pinpointing

Pinpointing in general compares the incoming speech to stored speech for the exercises in the system. To test the results of pinpointing individual phone errors, we trained on eight exercises designed to teach four sounds of English that are problems common to speakers of many different mother tongues, IH as in "bit", S as in "sit", T as in "tie", V as in "vine". For children's voices, we need to adapt our system to use children's voices as reference voices for the pinpointing, and then check the reasonableness of the pinpointer's classification of the student's voice. For this we had ten *other* young native speakers of English read the exercises we wanted to test on.

The eight exercises are already used in the adult system. We then needed to record non-native speakers reading the same content to determine the precision of both the adult and children's system.

4. THE TEST PARADIGM

We compared different conditions (adult system with adult and children's data, children's system) to determine whether changing data and number of tied states alone would give the child system as good pinpointing precision as the adult one.

4.1 Recording native and non-native adults and children for the comparison test

Sixteen non-native adults and fourteen non-native children read the exercises mentioned above. The target sounds appear in a variety of word positions and have a differing phonetic context. The 58 different utterances per speaker break down as: 19 "IH" utterances, 14 "S", 14 "T" and 11 "V". An utterance consists of one word or a contrasting two-word minimal pair containing an average of 5 phones ("seat sit"). There were a total of 4079 phone tokens for the adults and 2854 for the children (not everyone finished). The following table describes the speakers:

Adults	**No.**	**Children**	**No.**
Russian	7	Russian	2
Mandarin	4	German	4
Cantonese	1	Speech Impediments	3
Japanese	2	Japanese	2
Korean	1	Korean	2
Hindi	1	Hindi	1

Table 1. Origins of non-native and speech-impaired speakers – adults and children

Individuals in the two groups of speakers (adults and children) were not matched exactly as to language of origin, but there was a diverse mix of native languages for each.

4.2 Human tutor annotation as a base of comparison

In order to assess results on each version of the adult and children's systems, a human tutor's judgment was used as "ground truth": what the non-native actually said correctly and incorrectly. Only one expert human tutor listened to the data and noted for each phone whether it was understandable, or whether, if she had the student in a class, she would stop and correct the phone (because it was not comprehensible). Given more time, we would have had three tutors perform this task. In earlier work [5] that assessed the effectiveness of our system in actual pronunciation learning, the tutor (hereafter called rater) was asked to rate each phone 1 or 2, as to whether it was wrong or right, respectively. However, when several raters were asked to rate the same speaker using this metric, the inter-rater results were extremely variable. The rating scale was then changed from a binary to three levels. The rater could decide the phone was 1: completely wrong; 2: a good effort, but not quite comprehensible; 3: correct, comprehensible. This three-tier ranking was used in the results below. More information on human rating may be found in [6].

4.3 The comparison statistics

First we wanted to see what the user actually experienced, that is, would the system show errors for children as precisely as for adults on just the phone that was targeted in each utterance (the "focus" phone). There are few phones in any given utterance that are corrected (the targets of the exercise, "IH" for example) - only one to five occurrences in any utterance. This focuses the user's attention on one (phonetic) goal at a time. However, for a more reliable result and a better idea of how the pinpointing performs over all sounds of English, we have performed a second set of analyses over all of the phones of all of the utterances in all of the IH, S, T and V exercises above.

The classifier creation methodology used here is: have a human rater classify all the phones for all subjects; have the pinpointer classify all these phones according to some parameter value settings, and calculate the Cohen Kappa value (see below) for that parameter setting. We calculated the Kappa value for both the cases of folding the phones with the human intermediate rating (#2) value in with either the set of points with classified as bad (rating 1) and with the set of points classified as good (rating #3). All the results below have the rating #2 points added to the rating where the resulting Kappa value is the best.

4.3 Cohen's Kappa statistics

We used Cohen's Kappa [1], [3]. This statistic gives us with a measure of how well the automatic pinpointing agrees with the human expert for the N phones in any experiment. It takes into account the proportion of agreement that would occur by chance, if the raters are statistically independent. We made that assumption, since the human expert and automatic pinpointing use very different methodologies for carrying out the rating. Given that assumption, Kappa provides a reasonable assessment of how reliably the judgement of one rater is predicted by that of another.

The Kappa coefficient is calculated on an NxN matrix with the ratings for each rater on one of the axes. Each cell of the matrix contains the number of times the 2 raters assigned the corresponding cell values to classification items. The Kappa coefficient is calculated by first calculating the number of cases where one might expect the diagonal cells to occur by chance. This value is then used with the values along the diagonal to calculate the coefficient. The Kappa coefficient has a range from –1.0 to 1.0 with larger values indicating better reliability. Generally a Kappa of > .70 is considered very satisfactory.

5. RESULTS

The results are displayed first only on target phones (the impression the user has while using the system), then on all phones (general system precision).

Results for all configurations tested are shown in Figure 1. This figure shows agreement between pinpointer and human rater (at bottom), false positives (when the system told the user he/she was right and the human rater said the token was wrong) and false negatives (where the system told the user something was wrong that the rater had called correct). In general, high agreement is the most desirable characteristic. Changes can be made to systems to achieve a desired balance between false positives and false negatives. That tradeoff differs according to the type of application.

The first column, Adult/Adult, shows the performance of the adult pinpointer with Open Source SPHINX trained on Wall Street Journal (adult) data. We always show the parameter setting that gives the best human/system agreement while minimizing false negatives (our experience shows that users walk away from systems that tell them they are wrong when they know that they are not). Letting a few errors through may make it easier to concentrate on just some of the errors first. The second column shows the result of pinpointing using an adult-trained system on children's data. Here agreement is much lower, proving the necessity of using children's speech (or, lacking that, adapting to female speech).

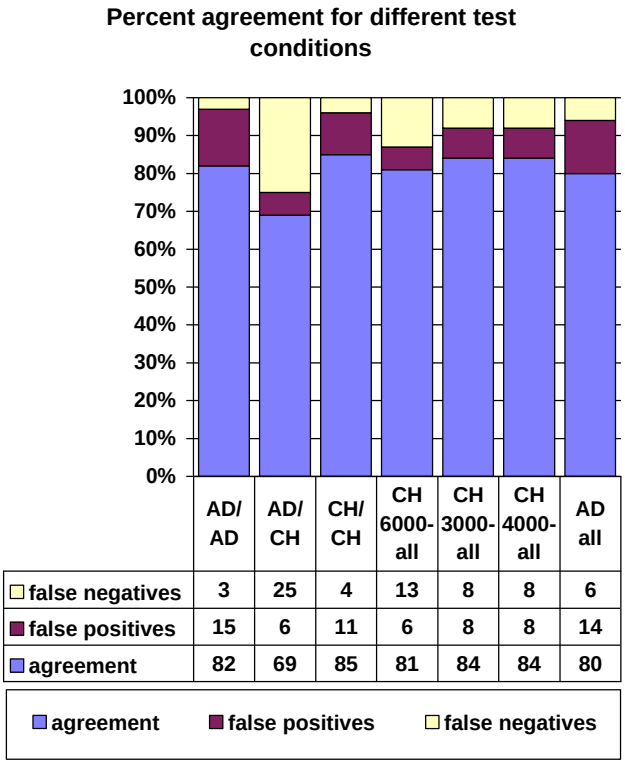


Figure 1. Percent agreement between the system and the human tutor. The first three columns show only the target phone in each utterance of an exercise. The four columns on the right show results over all phones all of the utterances. Note that we have merged the #2 ratings to the more favorable position since automatic pinpointing still makes a binary decision. Future versions of the pinpointer may include a "softer" decision. There were relatively few 2's – about 10% of all answers. (AD = adult, CH = child)

The third column represents the children's recognizer (4000 tied states). Agreement here is better than for the adult system.

Conditions	Kappa
Adult/ Adult	0.47
Adult/ Child	0.32
Child/ Child	0.35
Child 6000-all	0.28
Child 3000-all	0.28
Child 4000-all	0.27
Adult – all	0.37

Table 2. Kappa values for all test conditions

Table 2 shows the Kappa values. They are relatively low and indicate disagreement between automatic pinpointing and the expert rater. This is offset by the fact that we can correctly mimic human performance 85% of the time while only

characterizing a good sound as an error 4% of the time. Integrating such numbers into the skill assessment of the automatic tutor, allows this tutor to correctly change its estimate of a student's skill level given data about the performance/non-performance of that skill.

We can now examine results over all phones since there were more tokens in this case (669 tokens for the target sounds only and 4079 tokens over all phones). The last column in Figure 1 shows the base result for the adult system. The children's 6000-tied state system is worse than the adult base: agreement is lower and false negatives are high. But results for the 3000- and 4000-tied state systems are better than results for the adult system.

Unfortunately Kappa does not take into account total overall errors, or that the two different types of errors have different importance. So, while Kappa for the adult recognizer on the child speech almost matches Kappa for the child one, these other statistics make the adult recognizer less desirable.

Fleiss' weighted Kappa [7] would allow us to appropriately see that the adult system was worse at the expense of assigning a numeric weight to each error. In the future we will use the weighted kappa, finding a principled method of assigning the weights, and accumulating all comparisons in a single value [4].

6. DISCUSSION

For the user's overall impression as well as over all phones, the children's pinpointing performs as well as the adult version.

14 hours of kids' training data is thus acceptable for our application. This reinforces results shown for sparse data situations. Since recording in schools is becoming more and more difficult, this result is promising for those who need children's recognition systems.

We could provide a finer, more detailed measure of the quantity of data needed: with little data, the traditional tradeoff is between collecting a lot of data from few speakers and a small amount of data from many speakers (data in this project is close to the latter option – there are about as many different speakers as in the Wall Street Journal corpus the adult recognizer was trained on). Deciding how many speakers vs. how much diverse linguistic/phonetic content per speaker must be based on sets of tests concerning comparable-sized applications. There should be, rather than one general amount of speech measured in time, a guideline as to *how many speakers/how much speech/how much text* is needed to have good quality recognition.

Recent unpublished work at CMU has shown that an increase of 3-4% in accuracy of recognition can be obtained for the adult system using gender-dependent models. This will be tried soon as another way to increase pinpointing accuracy.

Decreasing the number of tied states from 6000 to either 4000 or 3000 has afforded good quality recognition. However decreasing again to 2000 decreased results considerably.

Since the main differences between the production of speech in children and adults reside in the length of the vocal cords and the length of the vocal tract, and since the main difference is in the latter with the descent of the vocal cords as a boy matures, we can presume that the speech of children, while very different from adults males, could be fairly close to female speech. We could therefore adapt to children's speech from adult female speech and thus bypass data collection. To our knowledge, there is no present system that performs as well using adaptation as using the speech from children, but we will, in the near future, try this method. This acoustic adaptation, while good for a read speech application, will not carry over to more spontaneous ones since word choice and grammatical structure also change.

The false positive/false negative tradeoff can be manipulated by our automatic (intelligent) tutor to help maintain its knowledge of the user's skills from the non-target phones in an exercise. The tutor only shows mistakes on the target phones of an exercise. It can use information from the non-target phones as evidence of skill achievement, and update its estimates of the non-target phone skills accordingly. If a student is failing to demonstrate some basic skill (like pronunciation of a vowel) that greatly affects intelligibility the tutor may choose to redirect the student to working on the more basic skill. This change of focus should occur only if necessary and only if the tutor is certain it is needed. Manipulating the recognition criteria so that false-negatives are minimized at the cost of possibly not changing focus will ensure that when a change of focus happens, it is indeed needed. We believe this bias toward maintaining focus is the right pedagogical approach Using the data from non-target phones should help the previous issue of how many times the user has to demonstrate a skill.

At this time, our best Kappa values for the adult test set are not high between the automatic pinpointer and the human rater. However, it is possible that with 2 or 3 human raters, agreement would be different. Thus we don't know how reasonable a Kappa of .32 is in an absolute sense; the Kappa did, however, give us a simple way to calculate a value that we could use as an optimization criterion; the Kappa values don't tell the entire story. We see that agreement with the human rater is high despite the low Kappa value.

7. A "YELLOW LIGHT"

After the initial version of this paper was written, another tests was carried out. It involved the human tutor's rating of 2. In results above, we had grouped the 2 rating with either 1 or 3, whichever case afforded the best results. In this test, we have treated a rating of 2 as a separate entity.

Since the human tutor was capable of using this intermediary notation reliably, it was felt that the user would benefit if the system could also furnish information. For the user, the intermediary rating is defined as a human tutor saying that "this is not quite it, but you are getting better, closer to the correct sound". Since pinpointing information is shown as the phone in question being shown in red (wrong) or green (correct), this appears as yellow, and results from the scores on the user's utterance being within a certain distance of the correct pronunciation. This distance (threshold) was determined as the setting that gave the best match to human rater results below.

Figure 2 below shows that adding this capability lowered the Kappa both for the adult and children's systems, mainly by classifying as intermediate results ones that had originally been classified as bad. However, even with the lowered Kappa, users perceive this change as positive. It gives a better intermediate classification to 40% of the false negatives for the children's version (using 3000 tied states) and 60% for the adult version, thus reducing perceived error.

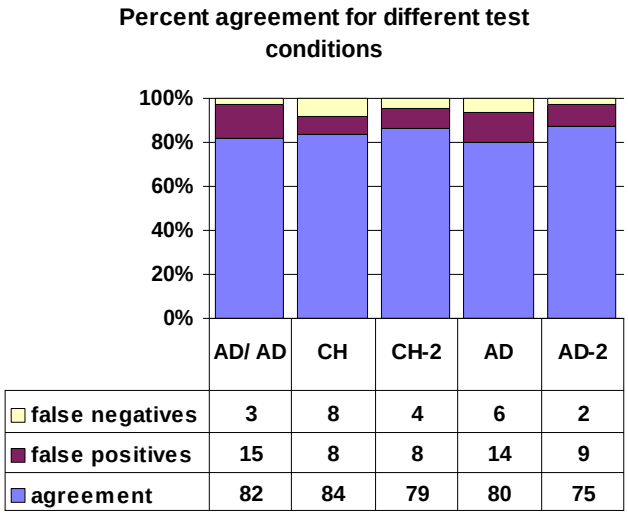


Figure 2. Changes in precision when a "yellow light" is added. The "- 2" indicates the versions where the "yellow light" was used. Note that the columns do not add up to 100% in the case of 2, since we are only noting false positives and false negatives as disagreements between 1s and 3s.

8. CONCLUSION

We have shown that retraining the recognizer and making few adjustments to the resulting system have given us a pinpointing system that is as good for children as it is for adults.

9. REFERENCES

[1] Cohen J., "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, 20, 37-46, 1960.

[2] Lee, K.F., Automatic Speech Recognition: The Development of the SPHINX SYSTEM, Kluwer Academic Publishers, Boston, 1989.

[3] Kraemer, H.C., "Kappa coefficient", In S. Kotz and N.L. Johnson (Eds.), Encyclopaedia of Statistical Sciences, John Wiley and Sons, New York, 1982.

[4] Landis, J., and Koch, G., "The measurement of observer agreement for categorical data", *Biometrics*, 33, 159-174, 1977.

[5] Mayfield Tomokiyo, L., Wang, L., Eskenazi, M, "An empirical study of the effectiveness of speech-recognition-based pronunciation training", in *Proc. ICSLP2000*, Beijing, 2000.

[6] Neumeyer, L., Franco, H., Abrash, V., Julia, L., Ronen, O., Bratt, H., Bing, J., Digalakis, V., Rypa, M., WebGrader: a multilingual pronunciation tool, in *Proc. ESCA- STILL98*, Marholmen, Sweden, 1998.

[7] Fleiss JL, Cohen, J., "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability", *Educational and Psychological Measurement*, 33, 613-619, 1973.

More details on the YOUTH database can be found at: http://www.carnegiespeech.com/products/YouthDB\youthDB.html