# Discriminative Model Combination and Language Model Selection in a Reading Tutor for Children

*Abdurrahman Samir, Jacques Duchateau, and Hugo Van hamme*

Katholieke Universiteit Leuven, ESAT
Kasteelpark Arenberg 10, 3001 Leuven, Belgium
e-mail: {Abdel-Rahman.Samir, Jacques.Duchateau, hugo.vanhamme} @esat.kuleuven.be

## Abstract

In this paper, we suggest the use of general acoustic and language models to deal with the mismatch between the training and testing data of a reading tutor for children. The testing data consist of isolated real and non-existing (pseudo) words, while the training data consist of continuous readings of Dutch sentences. General acoustic (e.g. context independent) and language models (e.g. bigram phone language models) are proposed as they implicitly better model the hesitant nature of the testing data. Discriminative model combination (DMC) is modified to provide different weights for different phones and was utilized to combine the new models into the baseline system. Combination of general acoustic and language models into the baseline system using DMC significantly lowers the system phone error rate, by 3.5% relative to the baseline system for the non-existing (pseudo) words.

**Index Terms**: reading assessment, ASR for children, Discriminative Model Combination (DMC)

## 1. Introduction

Automatic speech recognition (ASR) technology is utilized for building reading tutors and interactive language learning applications [1] [2]. The SPACE project[1] aims at utilizing ASR technology to provide schools, teachers and parents in Flanders with a reading tutor that targets children aged 6 to 10 years.

One of the challenges that face reading tutors targeting this age range is the hesitant nature of the speech, which results in producing unpredictable pauses and mispronunciations. Because the ultimate goal of such applications is to detect reading errors, analyze them and give corrective feedback, normal lexical trees that are used in the state-of-the-art LVCSR systems cannot be used anymore without modifications. One proposed solution [2] for such problem is to use subword units (e.g. syllables) in the decoding step.

In SPACE, a two pass system architecture is utilized [3] in which a phone lattice is generated during the first pass of the decoder (the task independent part) using general acoustic and phone language models only. Then the task-specific information is added during the second pass of the decoder in the form of a finite state transducer (FST) containing the correct phonetic transcription of the words along with garbage loops and garbage paths to account for pronunciations other than the correct one.

To assess the children reading proficiency level, three different categories of tasks are used: real words tasks, non-existing (or 'pseudo') words tasks and story reading tasks. This paper focuses on word reading tasks (both using real words and pseudo words). Using pseudo words in evaluating reading proficiency adds another challenge for the recognizer because they are not Dutch words (although they are constructed to resemble the real Dutch words). The pseudo words tasks contain new phone sequences which are different from those appearing in real Dutch words. In this work, we propose using context independent (CI) acoustic models along with context dependent (CD) ones because some triphones (from disfluent speech or pseudo words) could be poorly trained or could be combined with other acoustically different triphones in the decision tree classification process, which was trained on fluent utterances of real words.

Discriminative model Combination (DMC) [4] is used to combine the CI and CD acoustic models. The minimum word error criterion is utilized to estimate the model weights. DMC is extended to provide different weights for different phones. This gives better results than combining both models using only one weight vector.

On the language modeling side, the hesitant speech and the pseudo words don't follow the Dutch language phonotactic constraints. This suggests decreasing the depth of the n-gram phone language model used in the first pass of the decoder to better model reading disfluencies and new phone sequences. For phone recognition task, the bigram phone language model provides lower error rates compared to the trigram phone language model for pseudo words. By combining all knowledge sources, a significantly lower phone error rate is presented.

The following sections are organized as follows. Descriptions of training and testing speech databases are given in the next section. Section 3 contains a review of the DMC technique and the way we have used it in this paper. Experiments for different acoustic and language models are presented in section 4. Finally, in section 5, conclusions and ideas for future work are given.

## 2. Description of databases

### 2.1. Acoustic model training database

Acoustic models used in this work (both CD and CI) are trained on a read speech database in Dutch, recorded at 16 kHz sampling rate. Children aged between 5 and 11 years read different sentences. The database consists of 22 hours of speech in total from 400 children, distributed equally by age[2].

[1] SPeech Algorithms for Clinical and Educational applications. Home page: http://www.esat.kuleuven.be/psi/spraak/projects/SPACE.

[2] One exception: there are only few 5 year old children as in Flanders, most children learn to read at age 6.

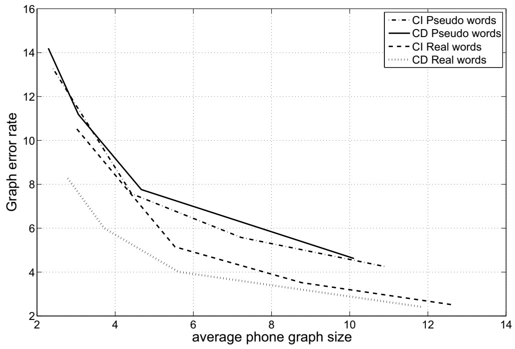

Figure 1: *The oracle phone error rate for CD and CI models vs. phone graph size.*

### 2.2. Language model training database

All phone language models used in this work are trained on a read speech corpus for Dutch news containing 65000 words. Forced alignment was performed on the corpus to select only one pronunciation in cases where there are multiple pronunciations for a word in the pronunciation dictionary.

### 2.3. Testing database

The CHOREC database [5] is used in all the experiments done in this paper. It contains read speech of 400 Dutch speaking elementary school children (6-12 years old) with or without reading difficulties.

For every child, a reading test battery was administered which contains real word reading tasks, pseudo word reading tasks, and story reading tasks. Both the real words and the pseudo words tasks contain three lists of respectively 40 1-syllable, 40 2-syllable and 40 3- or 4-syllable real words or pseudo words. Spoken utterances are provided along with their reference phonetic transcription. The phonetic transcription of utterances with errors, miscues, hesitations are manually verified.

In this work, only tasks with 2-syllable words (both real words and pseudo words) tasks are used. Recordings for 59 children (2320 real words, 2240 pseudo words) are used to train the DMC weights. While recordings for another group of 55 children (1680 real words, 1680 pseudo words) are used for testing. The rest of the database is kept for future research.

All used recordings are from children attending regular schools, i.e. pupils attending schools for children with reading or learning disorders were not included. However,the reading proficiency of our test children varies over a wide range.

## 3. Discriminative model combination

Discriminative model combination (DMC) [4] aims at integration of all given streams (or models) into one log-linear posterior probability distribution. Training of combination weights is done by minimizing the word error rate $E(\Lambda)$. Where,

$$E(\Lambda) = \sum_{n=1}^{N} \sum_{k \neq k_o} L(k, k_o) S(k, n, \Lambda) \qquad (1)$$

$L(k, k_o)$ is the edit distance between the $k^{th}$ hypothesis and the correct one $k_o$. $S(k, n, \Lambda)$ is a smoothed indicator function which is defined as,

$$S(k, n, \Lambda) = \frac{P_\Lambda(k \mid x_n)^\eta}{\sum_{k'} P_\Lambda(k' \mid x_n)^\eta} \qquad (2)$$

where $\eta$ is the smoothing constant and $P_\Lambda(k \mid x_n)$ is the posterior probability of the $k^{th}$ hypothesis given the weight vector $\Lambda$ for the feature vector $x_n$ of the $n^{th}$ training utterance.

To achieve the optimal weight vector $\Lambda$, the word error rate $E(\Lambda)$ is minimized using an iterative gradient descent algorithm.

In our implementation of the DMC technique, a phone graph is used, rather than an N-best list, so the rivals for the correct hypothesis at a certain frame are those competing arcs that have not been pruned out during the first pass of the decoder. Furthermore, a normalization over the number of frames $F_n$ in each utterance is performed. So $E(\Lambda)$ will be,

$$E(\Lambda) = \sum_{n=1}^{N} \frac{1}{F_n} \sum_{f=1}^{F_n} \sum_{k \neq k_o} L(k, k_o) S(k, f, \Lambda) \qquad (3)$$

where $L(.)$ is either 1 or 0. Up till now, only one weight per stream is estimated. To fully utilize the second stream, one weight per phone could be estimated. So the indicator function was modified to enable weight vectors for all phones to be jointly estimated. The $S$ function was modified to be,

$$S(k, f, \Lambda) = \frac{P_{\Lambda_k}(k \mid f)^\eta}{\sum_{k'} P_{\Lambda_{k'}}(k' \mid f)^\eta} \qquad (4)$$

where $\Lambda$ is a matrix that contains weight vectors $\Lambda_k$ for each phone $k$. The modified weights are normalized in every iteration of the algorithm so as to sum to one.

## 4. Experiments

### 4.1. System performance metrics

Two performance metrics are used in this paper to assess the system performance. The graph error rate (GER) is used to check the percentage of phones that are deleted, substituted, or inserted on a phone graph when searching for the best match for the database reference phonetic transcription (which contains errors, hesitations, etc...) of the spoken utterance. Each graph contains the first pass decoding result of the child's response for only one word. All graphs that are presented in the experiments are generated using the trigram phone language model during the first pass of the decoder.

The graph density is determined by decoder parameters: the beam width relative to the best hypothesis, and the maximum allowable number of competing hypotheses at a certain frame. We use the average number of distinct phones at a given frame as a measure for graph size. High quality phone graphs are required during the first pass of the decoder as they are the inputs to the second pass which cannot recover from errors made by the first pass.

The second performance metric used is the isolated phone recognition (PER). First, the utterance is aligned to its reference transcription using the Viterbi algorithm. A tight search beam was imposed during the Viterbi search to remove utterances that contain transcription errors. Segment recognition is done by ranking hypotheses based on their segment scores: the total segment likelihood in case of acoustic model assessment or the combined acoustic and language model likelihoods in

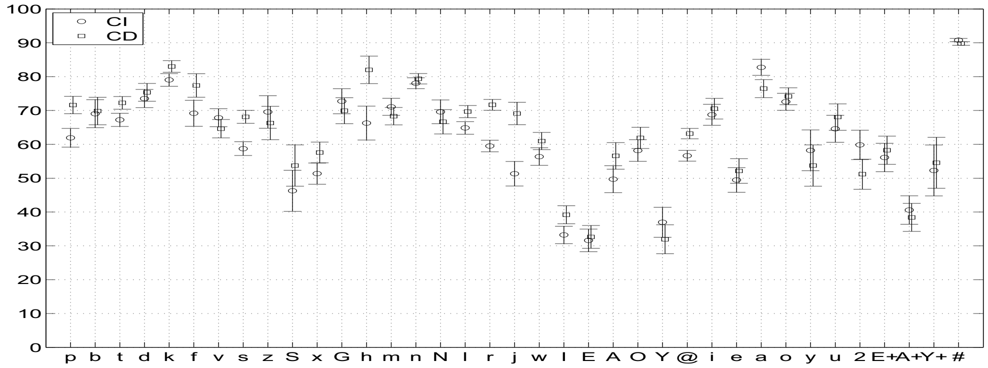

Figure 2: *phone accuracy for CD and CI acoustic models in real words with error bars of one standard deviation.*

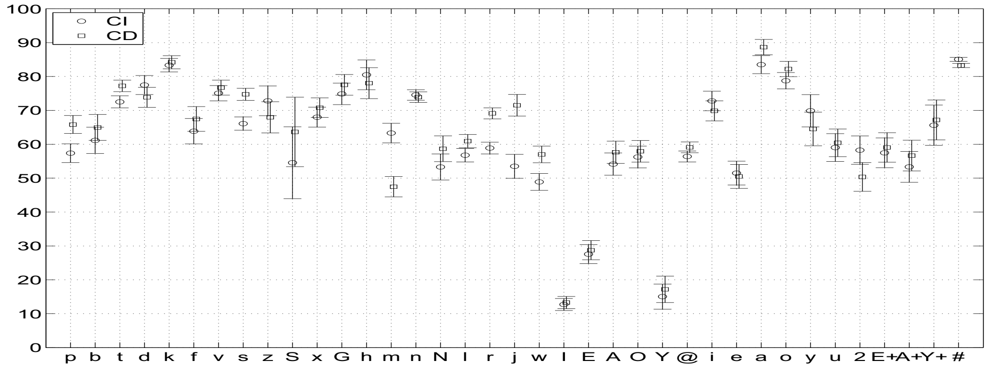

Figure 3: *phone accuracy for CD and CI acoustic models in pseudo words with error bars of one standard deviation.*

case of language model assessment. Combination of the acoustic model and the language model scores is done by using a fixed weight (the same weight is used for all experiments) that is tuned on a part of the training data. In cases where the segment context is needed to deduce the segment score, the correct context is provided so as to limit the search to the maximum number of monophones in the system.

Table 1: *Segment PER for CI and CD models and the combined systems.*

| | Real words | Pseudo words |
|---|---|---|
| CI | 30.7% | 32.3% |
| CD | 27.8% | 30.6% |
| single weight | 28.2% | 30.5% |
| weight per phone | 27.9% | 30.1% |
| oracle model per phone | 27.3% | 29.7% |

### 4.2. Acoustic models experiments

Segment recognition, as described in the previous subsection, is conducted on the real words and pseudo words tasks using Context Independent (CI) and Context Dependent (CD) acoustic models only (on phone language model is used in this subsection). Table 1 shows that the CI model has higher PER than the CD model for both the real words and pseudo words tasks. But the performance gab between the CI and the CD models are smaller in case of pseudo words task.

The pseudo words task differs from the training database because the pseudo words are not real Dutch words, although they are formed to resemble the real Dutch words. So it is expected to find triphones in the testing data that are not existing in the training data, which is not the case of the real words task.

The previous result is also supported by figure 1 which present graph error rates (GER) versus average graph size for the real words task and the pseudo words tasks. The CI model demonstrates higher capability of producing graphs with low error rates compared to the CD model in case of the pseudo words task, so combining both models to improve the system accuracy is worth exploring. Discriminative model combination (DMC) come to the play as a tool that is capable of estimating combination weights of multiple streams, in our case the CI and the CD acoustic models. The combined system, using one weight for each model, performs almost the same as the CD models in case of the pseudo words task and worse in case of the real words task. To combine the CI and the CD models in a more effective way, different weights per phone are estimated using the DMC. The new system is better than the CD models case by 0.5% absolute. These results are presented in Table 1.

The best acoustic model for a phone (either the CI or the CD model) is the one which gives lower PER for this phone. We define the oracle model as the one which picks the best acoustic model for each phone. Figure 2 and figure 3 show phone error rates per phone for the real words and the pseudo words tasks on the testing database. It is clear that the number of phones that tend to have better PER using the CI models in case of the pseudo words task is higher than the case of the real words task. Unfortunately, the phones that have better CI PER in case of CI models do not form a clear phonetically inspired class.

It is shown in table 1 that in case of the pseudo words task, the PER of the combined system, which uses both CD and CI acoustic model with different weights per phone, lays in the middle between the PER of the CD model (the baseline) and the PER of the oracle model. The oracle PER cannot be achieved in practice because if the CI model is better for a given phone, this assumes that all other competitors are modeled by the CI model. Likewise, for another phone, the CD model is better assuming that all other phones are modeled by the CD model. This assumption is not true when processing a phone lattice, as different phones have different model weights.

Table 2: *PER results using different phone language models for real words task.*

| | 0-gram | 1-gram | 2-gram | 3-gram |
|---|---|---|---|---|
| CI | 30.7% | 29.3% | 28.0% | 28.0% |
| CD | 27.8% | 27.2% | 25.0% | 25.2% |

Table 3: *PER results using different phone language models for pseudo words task.*

| | 0-gram | 1-gram | 2-gram | 3-gram |
|---|---|---|---|---|
| CI | 32.3% | 30.5% | 29.8% | 30.4% |
| CD | 30.6% | 29.0% | 28.3% | **29.0**% |

Table 4: *Perplexity of the 2-syllables real words and pseudo words tasks using different phone language models.*

| | 1-gram | 2-gram | 3-gram |
|---|---|---|---|
| real words task | 34.8 | 25.7 | 25.6 |
| pseudo words task | 36.2 | 29.9 | 39.8 |

### 4.3. Language models experiments

Much like the general acoustic model better models the pseudo words, general language models also provide better modeling of pseudo words. Tables 2 and 3 show the recognition results when using 0-gram, 1-gram, 2-gram, and 3-gram phone language models combined with different acoustic models for the real and the pseudo words tasks respectively. The bigram phone language model has the lowest error rates. For the real words task the 3-gram phone language model has comparable performance.

By reducing the depth of the n-gram, the language model becomes more powerful in modelling unexpected phone sequences both in the pseudo words and those introduced by the children's hesitations. Unpredicted pauses during reading and mispronunciations introduce new phone sequences which are unseen in the training data. Table 4 supports these PER results for different depths of n-gram by showing the perplexities of the 2-syllables real words and pseudo words tasks using these language models. The bi-gram phone language model gets the lowest perplexity in case of the pseudo words task.

Table 5: *PER results using the combined system with different language models.*

| | 0-gram | 1-gram | 2-gram | 3-gram |
|---|---|---|---|---|
| comb real | 27.9% | 27.0% | 25.6% | 25.7% |
| comb pseudo | 30.1% | 28.6% | **28.0**% | 28.4% |

### 4.4. Combined acoustic and language models

By combining all the knowledge sources together, Table 5 shows that the combined system, for the pseudo words task, using bi-gram language models (shown in bold in table 5) has a lower PER than the baseline system (shown in bold in table 3) by 1% absolute (3.5% relative). The difference between these two classifiers is statistically significant at the 5% level. No improvement is found in case of the real words tasks. One thing to note is that the differences in accuracy between different phone language models and the 0-gram case is small, which suggests further refinement of the phone language models.

## 5. Conclusions

We have combined general acoustic and language models into the phone recognition task of children's read speech. Discriminative model combination (DMC) was extended to provide different weights per phone and used to estimate the combination weights for the Context Independent (CI) and the Context Dependent (CD) acoustic models. The bigram phone language model was tested versus the trigram. Significantly lower phone error rate (3.5% relative to the baseline) was found, in case of the pseudo words by using the combined CI and CD models along with the bigram phone language model. Applying the best system configuration into the first pass of the decoder and further refinement of the phone language model are in our future work.

## 6. Acknowledgements

The research in this paper was supported by the IWT project SPACE (sbo/040102): SPeech Algorithms for Clinical and Educational applications, home page: http://www.esat.kuleuven.be/psi/spraak/projects/SPACE.

## 7. References

[1] Beck, J. E., Jia, P., and Mostow, J. "Automatically assessing oral reading fluency in a computer tutor that listens." Technology, Instruction, Cognition and Learning, vol. 1, pp. 61-81, 2004.

[2] Hagen, A., Pellom, B., and Cole, R., "Highly accurate children's speech recognition for interactive reading tutors using subword units", speech communication, vol. 49, pp. 861-873, 2007.

[3] Duchateau, J., Wigham, M., Demuynck, k., and Van hamme, H., "A flexible recognizer architecture in a reading tutor for children," in Proc. ITRW on Speech Recognition and Intrinsic Variation, Toulouse, France, May 2006, pp. 59?64.

[4] Beyerlein, P., "Discriminative model combination." Proceedings ICASSP, pp 481-484, 1998.

[5] Cleuren, L., Duchateau, J., Ghesquire, P. and Van hamme, H.,"Children's Oral Reading Corpus (CHOREC): Description and Assessment of Annotator Agreement", In LREC 2008, Marrakech, Morocco.