# TRAINING OF HMMS FOR PRONUNCIATION ERROR DETECTION – CROSSLINGUAL BOOTSTRAPPING VS. FLATSTART TRAINING

*Michael Beilig*[1], *Diane Hirschfeld*[2], *Oliver Jokisch*[3], *Uwe Koloska*[4]

[1,2,4]*voice INTER connect GmbH*

[3]*Institut für Akustik und Sprachkommunikation, TU Dresden*

beilig@voiceinterconnect.de

**Abstract:** This paper presents an evaluation study of hidden markov model (HMM) training as part of the ongoing EURONOUNCE project that intends an Intelligent Language Tutoring System (ILTS). The Core functionality of the system is a computer assisted pronunciation trainer (CAPT), integrating audio–visual feedback based on a pronunciation error detection on the phone–level. In a first step the application will concentrate on Slavonic languages and German. Most of current approaches solely use hidden markov models (HMMs) of the language to learn (target language models). One characteristic of our approach is to additionally use source language as well as intermediate acoustical models to detect and rate pronunciation errors. The other specific is that our methods to achieve this goal are directly based on knowledge of linguistic experts and experiences of pronunciation training. So the extended acoustic model inventory is used by an expert to formulate pronunciation error hypotheses for each of the training utterances strongly related to the source–target language (L1–L2) pair just as to specific learning levels. As an initial study a cross language bootstrapping approach for German as source- and Polish as target–language was implemented to train a native Polish monophone model set. By this way, the authors approximate the break–even amount of data, at which the cross language bootstrapping can outperform the flatstart training procedure.

**Kurzfassung:** In diesem Beitrag werden Teilergebnisse zum Training von Hidden Markov Modellen präsentiert, welche während der Arbeit im EURONOUNCE Projekt erzielt wurden. Hauptziel von Euronounce ist die Entwicklung eines computergestützten intelligenten Sprachlernsystems (ILTS) für das Aussprachetraining von slawischen Sprachen und Deutsch. Auf der Basis einer phonemgenauen Detektion von Aussprachefehlern sollen dem Lernenden durch audio–visuelle Rückkopplung der aktuelle Lernerfolg sowie Hinweise zur Fehlerkorrektur zur Verfügung gestellt werden. Ein Großteil derzeitiger Ansätze verwendet ausschließlich Hidden Markov Modelle (HMMs) der zu erlernenden Sprache (Zielsprach–Modelle). Im angestrebten Ansatz wird dieses Inventar um Quellsprach- und lernfortschrittsspezifische Zwischen–Modelle (Intermediate–Models) erweitert. Eine weiteres Charakteristikum des Ansatzes ist die direkte Nutzung des Wissens von Sprachexperten und deren Erfahrung im Fremdsprachenunterricht. In der Realisierung bedeutet dies, dass Sprachexperten für jede Übungsäußerung konkrete Fehlerhypothesen mit Bezug auf das L1–L2 Paar sowie die Stufe des Lernfortschrittes festlegen. Erste Trainingsergebnisse für das Training von Monophon–Modellen für das Sprachpaar Deutsch–Polnisch unter Berücksichtigung der geringen Datenmengen wurden durchgeführt. Evaluiert wurden Ansätze des überkreuzsprachlichen Transfers von akustischen Modellen.

# 1 Introduction

The methods of computer–assisted language learning and intelligent language tutoring systems (ILTS) play an increasing role in the second language education. The ILTS system, used in the EURONOUNCE project [2], provides speech signal analysis functions on the users' speech input which allows the user to compare his pronunciation with the tutors' one and to receive selective information about potential improvements of articulation. The speech signal processing involves speech recognition and feedback technologies using signal analysis. The baseline platform AzAR (German acronym for 'automat for accent reduction') was developed in preceding projects 2005–2007 [3][4]. The core function is based on different phonetic–phonologic and prosodic distance measures, involving typical crosslingual influences from a native source language on the target language. It leads to the marking of mispronounced phones within a spoken utterance (see Fig. 1) using a coloured scale from red ("very bad") to green ("very good"). The system is using confidence measures from a HMM based speech recognizer. The didactic content follows conventional lessons for the pronunciation training with contrastive exercises, insertion tests, etc. The software was originally developed for Russian migrants learning German. In the euronounce project, the chosen sets of source (native) languages L1 and the target (taught) languages L2 include widely–used languages like German and Russian and national languages which are less taught as foreign languages, like Polish, Czech and Slovak. The project supports teaching and private studies of languages in neighbor countries, e. g. using e–learning infrastructure and computer–based language courses. The project started November


**Figure 1** - AzAR trainer application with visual feedback of detected and rated pronunviation errors

2007 and is going to be finished in October 2009. Euronounce is also intended to build an interdisciplinary and multinational network of teachers, linguists, phoneticians, speech technologists and experts for dissemination. The research and development has focused on Slavonic languages so far and crosslingual effects on German or vice versa. The development consortium consists of TU Dresden and voice INTER connect GmbH (Germany), Adam Mickiewicz University in Poznan (Poland), Slovak Academy of Sciences in Bratislava and Russian Academy of Sciences in St. Petersburg. The technically oriented tasks of these partners consider following basic issues:

- Specification of speech data, algorithms and tools;
- Recording and annotation of multilingual speech databases for L2 tutoring and for ASR;
- Data analysis and speech signal processing;
- Integration into tutoring and courseware systems;
- Technical support for evaluation and dissemination.

Acknowledged partners from language education, e. g. the Goethe–Institut, evaluate the systems concept in real language courses.

## 2 Concepts of pronunciation training

To successfully install computer assisted pronunciation training, several important prerequisites such as robust and detailed error analysis and selective feedback must be given. Some CAPT approaches have focused on ASR based global score of whole sentences or utterances, but for selective feedback and accelerated learning, an robust error detection on segmental level is required. Therefore most CAPT systems and related researches deal with detection and scoring on the phone segment level [5]. In the field of pronunciation error detection and analysis there are various approaches. While some researches deal with classification methods as LDA or Decision Trees [8], the majority is based on likelihood related features using techniques of automatic speech recognition including HMM–based acoustic modeling. The sequence of subprocesses is similar on most systems. In a first step a forced alignment recognition is applied based on the canonical transcription of the utterance using acoustical models of the target language. In a next step some a pronunciation score is determined using based on the segmentation marks and the rusulting likelihood. Common approaches to obtain this score are based on likelihood ratio and posterior probability. A well known score is the “Goodness of Pronunciation” (GOP) [9] that corresponds to a frame–normalized ratio of the likelihoods from forced alignment and an additionally performed free–phone classification. For the task of speech recognition each phone model is trained with a broad range of allophones to cover most inter- and intraspeaker differences. But to detect pronunciation errors more information about the utterance is needed and can be provided by statistical evaluation if there is enough speech data to be statistically relevant. Because for L2 data it is very difficult to collect such a great amount of adequately annotated speech data, other sources of knowledge has to be used. In the ISLE project [5] there is some rule based approached to expand the canonical transcription by mispronunciation hypotheses using acoustical models of the target language and perform forced recognition on the resulting lattice. Cause on principle native models are not able to cover the full range of appearing pronunciation errors, there have been approaches to explicitly train mispronunciation models [1].

## 3 Detection and rating of pronunciation errors in Euronounce

Former researches has been taken into consideration in AzAR [3] that provides the pronunciation training module in the Euronounce project. As mentioned, to successfully apply methods of statistical learning for our purposes a number of expert annotated and language pair specific speech data would be required, that to produce is infeasible for the present. On the other hand it is difficult to describe formal rules that sufficiently cover all potential errors and however reduces them to those that are most likely observed for a concrete training utterance. The characteristic of the AzAR concept is therefore to overcome these shortcomings by delegating the problems to the knowledge of linguistic experts. Based on their technical knowledge and not least on their practical experiences in pronunciation training the following fundamental tasks are transfered to them:

- construction of specially crafted training sentences with strong focus on L1–L2 specific pronunciation errors
- constitution of error hypotheses with respect to the different learning levels
- definition of specific intermediate phonetic entities appearing on different learning level
- provision of systematic audio–visual feedback depending on the occurred pronunciation error

The constituted error hypotheses composed of source–language, target–language and intermediate models will finally be the input of the ASR system to decide what sequence was most likely realized. All entities of the model inventory have to be trained afore. The mentioned problem of sparse training data especially relating to the intermediate models is intended to be resolved by applying adapting techniques as MLLR or MAP to the most similar model of the target or source language. Therefore in a first step model sets of all intended source and target languages have to be trained. So the current work presented is an evaluation of a cross–lingual approach to train a Polish monophone model set.

## 4 Experiments

As an initial study for the integration of Slavonic languages to the intended detection of mispronunciations on the segment–level, a native Polish monophone model set has to be trained. The available amount of native Polish speech data (described in 4.1) however does not securely fulfill the quantity requirements of a flatstart training procedure, to lead to sufficient models. A bootstrapping by native Polish data was not feasible cause no manual labeled data was available at this moment. To overcome this possibly insufficient amount of data a cross language bootstrapping approach [6] was taken into consideration. So the object of research was to evaluate the flatstart verses the cross language bootstrapping approach to train a native Polish monophone model set using a given amount of Polish training data and an already trained German monophone model set for bootstrapping. The evaluation was realized by a monophone–loop recognition. All experiments were performed under the usage of the HTK–Software [7].

### 4.1 Training data

For the training of Polish monophone models speech data recorded and annotated at the Adam–Mickiewicz–University (AMU) were used. This database contains Polish and German text corpora, both read by Polish and German native speaker, so that native and nonnative speech

data of both languages are available. For the current experiments, only Polish language speech data read by native speaker were applied. Phonetic transcriptions are available for all data. The training data holds the following attributes:

- phonetically balanced and rich sentences
- number of speaker: 14
- total number of sentences: 6336

For training and evaluation purposes the data–corpus was split complementary, so that the the evaluation corpus holds one male and female speaker and a total number of 1039 sentences.

### 4.2 Training methods

For flatstart and cross language bootstrapping, following feature extraction and model types were used:

- 12 MFCC's + log energy + delta + delta delta
- continuous densities, diagonal covariances

#### 4.2.1 Flatstart training

In a first step prototype models were created to define the structure of any model to train. The mean and variance values of the Polish monophone prototypes were initialized by the value of global means and variances calculated on the complete Polish training data. The training process was realized by five times iterative training using the embedded Baum–Welch Algorithm for re–estimation [7] (Iter1 to Iter5), followed by increasing the number of mixtures by two whereas two training iterations were performed after splitting (Mix2 – Mix16).

#### 4.2.2 Cross lingual bootstrapping

Single gaussian German monophone models trained on a large speech corpus were used as seed models. The mapping of Polish to German monophones (see 1) was phonetically motivated and relates to the IPA scheme. In the function of seed models, the mapped German model was used to initialize the mean, variance and transition matrix values of the according Polish model. After initialization the Polish models were trained in analogy to the training processes performed in the flatstart approach.

**Table 1** - phoneme mapping table

| Pol | Ger | Pol | Ger | Pol | Ger | Pol | Ger | Pol | Ger | Pol | Ger |
|---|---|---|---|---|---|---|---|---|---|---|---|
| i | i: | p | p | J | g | s' | S | t^s' | tS | r | r |
| e | E | b | b | f | f | z' | z | d^z' | ts | w | v |
| a | a | t | t | v | v | x | x | m | m | j | j |
| o | O | d | d | s | s | t^s | ts | n | n | w~ | v |
| u | u: | k | k | z | Z | d^z | ts | n' | n | j~ | j |
| y | Y | g | g | S | S | t^S | tS | N | N | | |
| e~ | E | c | k | Z | Z | d^Z | tS | l | l | | |

#### 4.2.3 Cross lingual transfer

Polish models were not actually trained in this approach, but all their values were substituted by the values of their German pendant of the according training level, before recognizing the Polish evaluation data. This is just to give an indication of the improvement in performance by using Polish speech data in the training process.

### 4.3 Results

Evaluation was performed by a monophone loop recognition on the evaluation data. To approximate the break–even amount of data, at which the cross language bootstrapping outperforms the flatstart training procedure, the complete training data was successively reduced.

- complete data: 14 speaker, 6336 sentences
- reduced data: 7 speaker, 2124 sentence
- less data: 7 speaker, 834 sentences
- sparse data: 4 speaker, 243 sentences

The recognition results (figure 2, 3, 4, 5) show, that for the available amount of Polish speech data there is no improvement in recognizer performance by training the models using the inspected cross language bootstrapping approach. Only for sparse data it could be shown, that initializing Polish monophone models with adequate German models results in a higher phone–correctness for the case of monophone loop recognition.

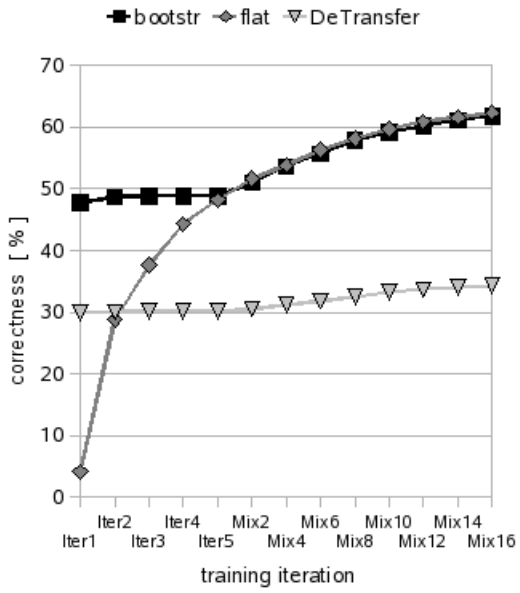


**Figure 2** - training on complete data

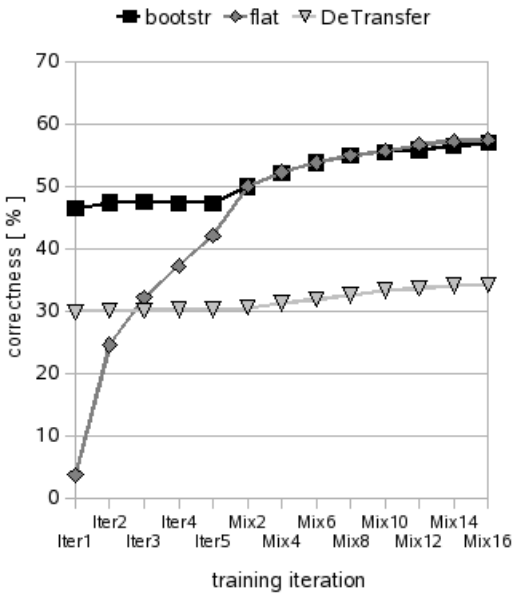


**Figure 3** - training on reduced data

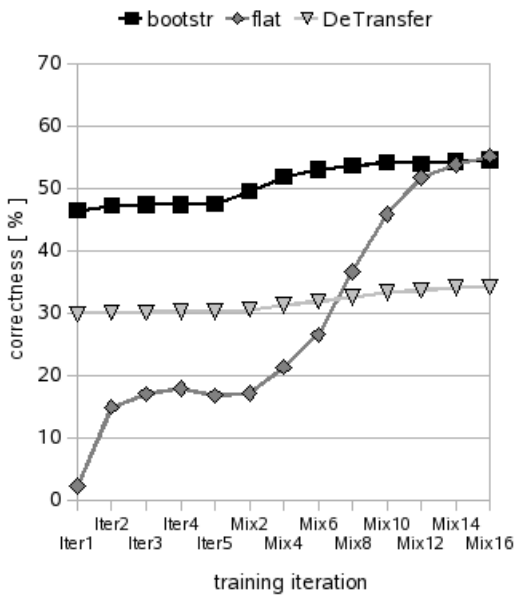


**Figure 4** - training on less data

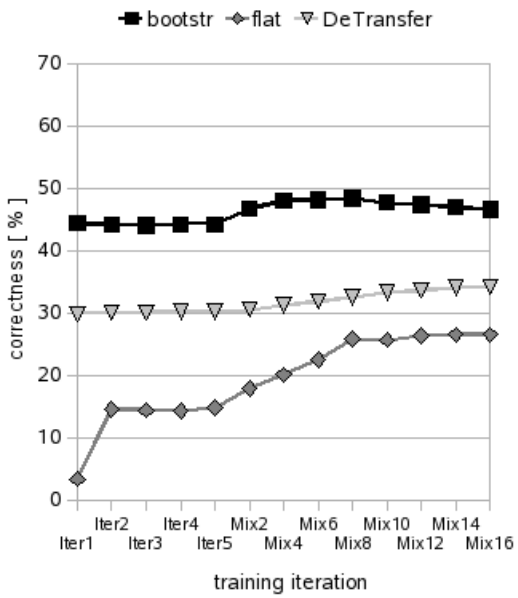


**Figure 5** - training on sparse data

## 5 Conclusions

The presented study was just the first step in realization of the intended approach and its integration into the pronunciation training system. In next steps training of the native models for the remaining languages Czech, Slovak and Russian will be executed. Further research will explore the generation of intermediate models based on the native models by evaluating different adapting techniques.

## 6 Acknowledgements

The authors would like to thank Agnieszka Wagner and Natalia Cylwik from University of Poznan, Poland, for preparing the Polish Database. This project has been funded with support from the European Commission within the Lifelong Learning Programme (project 135379-LLP-1-2007-1-DE-KA2-KA2MP). This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein. The project homepage is located at: `http://www.euronounce.net`.

## References

[1] FRANCO, H., L. NEUMEYER, M. RAMOS and H. BRATT: *Automatic Detection Of Phone-Level Mispronunciation For Language Learning*. In *EUROSPEECH'99*, pp. 851–854, 1999.

[2] JOKISCH, O., R. JÄCKEL, M. RUSKO, G. DEMENKO, N. CYLWIK, A. RONZHIN, D. HIRSCHFELD, U. KOLOSKA, L. HANISCH and R. HOFFMANN: *Euronounce Project –*

*an intelligent language tutoring system with multimodal feedback functions, roadmap and specification*. In *Proc. 19th ESSV 2008*, pp. 116–123, Frankfurt/M., September 2008.

[3] JOKISCH, O., U. KOLOSKA, D. HIRSCHFELD and R. HOFFMANN: *Pronunciation learning and foreign accent reduction by an audiovisual feedback system*. In *Proc. 1st Intern. Conf. on Affective Computing and Intelligent Interaction (ACII)*, pp. 419–425, Beijing, October 2005.

[4] JÄCKEL, R., O. JOKISCH and R. HOFFMANN: *Evaluation of the Speaker Proficiency in a Pronunciation Tutoring System*. In *Proc. 12th Intern. Conf. Speech and Computer (SPECOM)*, pp. 772–777, Moscow, October 2007.

[5] MENZEL, W., D. HERRON, P. BONAVENTURA and R. MORTON: *Automatic detection and correction of non-native English pronunciations*. In *Proceedings of InSTILL 2000*, 2000.

[6] SCHULTZ, T.: *Multilinguale Spracherkennung: Kombination akustischer Modelle zur Portierung auf neue Sprachen*. PhD thesis, University Karlsruhe, 2000.

[7] STEVE, YOUNG, E. A.: *The HTK Book (for HTK Version 3.4)*. Microsoft Corporation, Cambridge University Engineering Department, 1995–1999, 2001–2006.

[8] TRUONG, K. P., A. NERI, F. D. WET, C. CUCCHIARINI and H. STRIK: *Automatic detection of frequent pronunciation errors made by L2-learners*. In *INTERSPEECH-2005*, pp. 1345–1348, 2005.

[9] WITT, S. and S. YOUNG: *Computer-assisted Pronunciation Teaching based on Automatic Speech Recognition*. In *Language Teaching and Language Technology*, pp. 25–35, 1997.