Cross-Language Experiment

Jakub ŠŤASTNÝ, Pavel SOVKA

Dept. of Circuit Theory, Czech Technical University, Technická 2, 166 27 Prague, Czech Republic

stastnj1@feld.cvut.cz, sovka@feld.cvut.cz

Abstract. *The contribution addresses the cross-language experiment. The aim was to test the possibility of the conversion French phoneme models into Czech ones. This model conversion uses the Hidden Markov Models (HMM) classification procedure. The first step consists of the iterative mapping of French models to Czech ones. The mapping is given by the analysis the confusion matrix. The second step is the Baum-Welch re-estimation resulting in the final models for Czech language. Despite of the differences between French and Czech languages the final recognition score reaches 64% for the phoneme recognition and 74% for digit recognition. Relatively low recognition accuracy is caused by the inadequate noise model.*

The experiences gained with the cross-language experiment were utilized for the classification of simple human body movements. The solution of this problem and results are described in the second part of this contribution under the title EEG Signals Classification-Introduction to the Problem.

Keywords

Speech recognition, phoneme model mapping, hidden Markov models, HTK.

1. Introduction

The intense development of information technologies, the conception of the digital (paper-less) office and the mobile communication boom caused equally fast development and consequent real-life usage of speech-classification systems. Among interesting applications the InfoCity can be found ([1] and [2]) - a unique information system in Liberec based on the directed dialogue between the user and the system. The system offers to the user menus by means of synthesized speech, the user chooses what he/she wants by saying his/her choice or enters data (presumed departure of the bus, etc.). Another interesting application is the telephone voice dialing - nowadays already well-mastered technology. Voice dialing allows the car driver to leave the hands on the steering wheel even during the call set-up - he/she doesn't have to dial manually - and thus it improves the road traffic safety. Voice dialing system must be quite robust, it is supposed to work in every imaginable environment - in a silent office (the phone in a manager's pocket), on the airport or in a running car at a speed of 150 kmph during ride (see e.g. [3]).

Hidden Markov Models have been typically used for the speech recognition nowadays. Their dominant usage is caused by the following advantages: *clear architecture* of the classifier allows the user to reliably found out what is the system trained for and *model design* may be closely related to the real physical character of modeled random process. The model constructed with taking the real system properties into account and properly trained in such a way can be used as a source for the deep statistical analysis of the classified random process. Moreover, the mapping to the physical reality allows the usage of *context information* for increasing the classification score. Last, but not least, the *ability of modeling* should be mentioned. The model is able to generate synthetic realizations of the underlying random process.

HMM-based systems are typically robust and they have a large ability of generalization - InfoCity is able to handle many speakers and voice colors; the voice dialing system must work more or less independently on the ambient noises.

Worth mentioning is the fact that the kernel of the model training is a hard approximation problem [4]. Nevertheless, the quite reliable algorithm known as the Baum-Welch re-estimation procedure exists. Detailed information about HMM can be obtained from references [5], [6], [7].

In this contribution we describe the cross-language experiment targeted to the examination of how successful phoneme models trained for one language can be converted to the phonemes suitable for another language. Such a cross-language task arises in the case when the conversion of any existing speech recognition system from one language to another one is needed. Then the designer has to modify the models and the grammar. Phoneme models usually have to be retrained from scratch. The simplification can be achieved by utilizing the cross-language approach, where the original phoneme models are assigned to similar phonemes in the target language and then usually retrained (bootstrapping). This process does not need training on the exhaustive database. One example of this process will be presented now - the conversion of Swiss French phoneme models to Czech phoneme models. For the phoneme modeling, classification and conversion Hidden Markov Models were used.

2. Cross-Language Experiment

As written above the task is the use of Swiss French phoneme models for the recognition of Czech language phonemes. This task requires the appropriate conversion of French phonemes into Czech ones. The conversion consists of two steps - the phonemes mapping and phonemes re-estimation. Two approaches can be used for the phonemes mapping. Firstly, the automatic approach when the information contained in the confusion matrix can be used. The second possible approach is rather a manual one when the linguistic description is used by a (human) expert. The latter approach can also give the apriori information for the automatic phoneme mapping. Both approaches used in this study will be described in details later. After the mapping of French phonemes to Czech ones evaluated on the training data set, the phonemes re-estimation is done on Czech sentences with the second subsequent evaluation.

2.1 Input Data

The training and testing parts of Czech data were carried out on three types of databases. All the data were sampled at the rate of 8 kHz, 16 bit linear quantization.

- Database Polyphone: The original French models were trained on 793 sentences uttered by 40 female and 40 male speakers from the Swiss French Polyphone telephone database [10], [12].
- Database D1: The database containing 80 Czech sentences (sport news and weather forecast) spoken by 15 speaker along with their phonetic transcription ("write, what you hear") was used for the phoneme conversion (mapping and re-estimation) and final testing. Of course, the phoneme conversion and testing used two disjunctive parts of the database.
- Database D2: The final testing with converted phonemes was evaluated using another set of 100 Czech sequences containing 492 digits spoken by other 15 speakers.

2.2 The Tool for Phoneme Modeling and Conversion -HMM

As stated above, Hidden Markov Models were used for the whole experiment. Let us specify the used approach in more details.

HMMs represent the paradigm for stochastic processes modeling. HMMs are twofold stochastic automaton (generators) each transition between states is probabilistic; each emitting state generates a random output. The distributions of output random variables are usually normal (Gaussian) ones.

HMM are typically described by the notation $\lambda = (A, B, \pi)$, where $A = a_{ij}$ stands for the transition probability matrix, a_{ij} is the probability of transition from the state i to the state j, $i, j = 1, 2, \ldots, N$.

N is the number of emitting states. Emitting states are numbered $1, \ldots, N$; non-emitting ones have numbers 0 and $N+1$; non-emitting states allow the easy connection of the models into strings for the continuous speech recognition.

$B = \{b_j(\boldsymbol{o})\}$ is the likelihood of output emission $\boldsymbol{o}$ in state j. Function $b_j(\boldsymbol{o})\}$ specifies the output random variable distribution. This can be of any kind and the output generated variable may be even discrete. Anyway, typically normal distribution is used.

π- models have specified the probability $\pi(n)$ of random process starting in the state $n$for each state.

One pass through the model $\lambda = (A, B, \pi)$ results in the generation of one realization of (studied) random process. The isolated word recognition (classification) is done by the inverse approach to the modeling. The likelihood $P(O|\lambda_i)$of random process generation by models λ_i, $i = 1$ to N is calculated for the sequence O gained by the speech signal parameterization.

Under the assumption the observable sequence O and model λ are given, following problems have to be solved.

- *Model construction*: HMM may have various architectures dependent on the claims on the system, training the structure and the character of the modeled random process [6].
- *Training*: in what a way to set model parameters to maximize $P(O|\lambda)$.
- *Sequence and recognition*: how to find a sequence of model states, which generates the given O and simultaneously is the most likely of all. This problem is in a close relation with the preceding problem of training.

Concerning the technical details, freely available toolkit implementing Hidden Markov Models called "HTK" was used [8].

Above stated problems and their solution will be now described in more details.

2.3 Model Construction

The design of the model and the suitable speech parameterization is the first step of the classifier construction. Owing to the character of speech signal the models of the structure named as "left-to-right, without skips" are often used (see Fig. 1). The number of the states is chosen on the basis of number of modeled random process stationary intervals. The number and the type of the random output distribution parameters are determined by the modeled process properties as well as the type of the output distribution itself. Linear prediction coefficients (LPC), Mel cepstrum (MFCC) and total signal power estimation are typically used as parameters. The first and the second order differences of the mentioned variables are often used, too (e.g. see [8] and [9]).

Data used in this experiment were parameterized using MFCC. Phoneme models (HMMs) for the recognition used three data streams; the parameter vector for each stream was built up of 13 parameters; the first stream processed 12 MFCC coefficients and signal energy estimation, the second stream contained 12 the first derivatives MFCC and energy signal differences, the third stream consisted of 12 the second derivatives MFCC and signal energy differences. Used phoneme models were composed of five states left-to-right without skips architecture mentioned above. Three emitting states of each phoneme used 20 mixed normal distributions for modeling non-gaussian underlying speech process. Because phonemes were classified, each phoneme had its own model with appropriate parameters of random process.

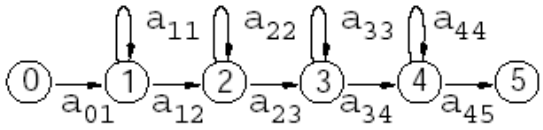

Fig. 1. The Markov model used for classification - schema. States 1-4 are emitting, 0 and 5 are non-emitting (connecting). The arrows display the possible transitions among states.

2.4 Training Problem

During the model training, matrices A and B were set in such a way to maximize likelihood $P(O|\lambda)$. To reach this one kind of EM algorithm (Baum-Welch re-estimation BWR) is used. The algorithm works iteratively, in two phases: *initialization* which gives the rough setting of model parameters and *re-estimation* which tunes both matrices A and B as precise as possible. Initialization step is a fast search of suitable initial conditions for the precise, but more time-consuming, re-estimation. This allows the re-estimation to give better results. Moreover, the proper choice of the number of steps for both algorithms is crucial for obtaining a good recognition score. Too small number of steps results in models not enough trained, too high number of steps causes limited system generalization ability. The determination of the optimal number of steps needed for the re-estimation applied on mapped models was one of important parts of the experiment (see next sections).

2.5 Sequence and Recognition Problem

The cornerstone of the recognition is the value $P(O|\lambda)$ computation problem. This can be interpreted as the evaluation of the likelihood that model λ has generated the observation vectors matrix O. The evaluation can be done in two ways.

1. By generating of all possible state sequences of the given length and properties. For each sequence the likelihood $P(O|\lambda)$ will be calculated and the total likelihood will be obtained by summing.
2. By finding the state sequence, which maximizes the likelihood O of generation.

The first way is used during system training - in BWR algorithm - because it provides results with higher precision than the second (utilized by in Viterbi decoder algorithm).

The HTK toolbox [5] offers two ways for the performance evaluation: the recognition score (rec) and recognition accuracy (acc). These two measures differ in the evaluation of the phoneme (or word) substitutions for pauses. Therefore when the noise model is inadequate the recognition accuracy is less than the recognition score. Thus the recognition accuracy gives the information how the noise model is matched to the environmental noise in the database used.

2.6 Continuous Speech Recognition vs. Isolated Word Recognition

A model chaining mechanism is necessary for the connected word recognition when the phoneme models are available. That's why each phoneme model has two non-emitting states for connecting with the previous and the next phoneme model into the chain. The result is a chain of phoneme models. The way of the phoneme models connection is described by a grammar. The grammar can be entered, for example, by the regular expression and the output of the classification is the string satisfying the given expression. During the connected word recognition a modified version of Viterbi algorithm is used - Viterbi beam search (for more see e.g. [7]).

2.7 Phoneme Mapping

At the beginning of the experiment it was necessary to make an initial mapping French o off phonemes on Czech ones according to the phonetic structure of both languages. The mapping was realized in two ways.

Heuristic mapping based on the expert's knowledge. A human expert decides for each Czech phoneme which French one resembles to it as close as possible. Subsequently, on the basis of this information a phoneme mapping is constructed. This step also includes the rough phoneme classification into basic groups (voiced, unvoiced and more detailed fricatives, consonants,...) used as the initial step improving the algorithmic mapping (see below). The heuristic mapping can serve also for the verification of results gained by algorithmic mapping.

Algorithmic mapping which utilizes the information contained in the normalized confusion matrix. Using this confusion matrix it can be stated which French phonemes can be unambiguous mapped to Czech ones and thus which French models shall have assigned to the appropriate phonemes. The recognition and subsequent mapping is repeated until mapping of all phonemes is satisfactory. Detailed description follows.

The algorithmic mapping proceeds as follows. Firstly the Czech grapheme-phoneme converter is applied to two selected sets of 91 Czech sentences recorded from broadcasting (2 male, 1 female speaker). One set is intended as learning, the other was used for the subsequent evaluation. After this, unconstrained phoneme recognition with French trained phonemes is performed on learning data. Then the confusion matrix is analyzed and French-Czech phoneme relations are built on the basis of this analysis.

The confusion matrix analysis can be summarized as follows.

1. Phoneme classification into basic groups (voiced, ...). This step simplifies the next analysis of the confusion table and ensures excluding the incorrect phoneme mapping.
2. Looking for maxima in the columns (rows) of the confusion table. Only the strongest maxims in the confusion table are used. These maxims yield French phonemes with unambiguously mapping to Czech phonemes (French phoneme model was often used for Czech phoneme classification during the experiment).
3. Transcription: the transcription amendment using the results of step 2 is performed and all steps (1-3) are repeated until the mapping is not completed for all phonemes.

Of course, not all phonemes can be unambiguously mapped to their counterparts. Possible alternatives must be always considered in step 2. All the steps are repeated until no further improvement of the recognition score can be done. The total number of phonemes used for mapping was 35 French phonemes to 43 Czech ones. Thus the different number of converted phonemes implies the ambiguous mapping. That is why the used algorithm was slightly more complicated than the one described in this text. Here we only outlined the basic steps of the mapping procedure used. The resulting mapping showed that the majority of Czech phonemes were assigned to their French similar counterparts (e.g. *p→pp, t→t, e→e*, but *H→r*). Results for mapped phonemes can be also seen from Tab. 1 (the first column).

2.8 Model Re-estimation

Based on gained information a new training was performed: the re-estimation of French phonemes with the Czech sentences. This procedure results in the increasing of the recognition score. French mapped phonemes were adapted to Czech using the learning set of 91 sentences. This was realized by applying Baum-Welch re-estimation procedure. During phoneme bootstrapping the dependency of the final recognition score on the number of BWR iterations was tested. As the BWR algorithm possesses good properties (strong iterative force and fast convergence) it was sufficient to perform only a few iterations (3-6). Table 2 summarizes the recognition score dependency on the number of iterations. Quite fast convergence rate of the system can be seen. The recognition score increases up to the third iteration. After the third iteration the score stays nearly constant (or it is slightly decreasing) but the models are loosing their generalization ability. This is typically caused by not sufficient amount of data given the number of iterations.

The retrained models were again tested and their performance was evaluated. This testing was done again on the 91 testing sentences as well as on the digit recognition database. The final results are summarized in Tab. 1 (the second column). By inspection it can be seen that if the expert's knowledge is used during the construction phase of mapping, the results are similar as for the algorithmic mapping.

2.9 Performance Evaluation

After the phoneme mapping and re-estimation the performance test was evaluated. The French context independent phoneme models with the found mapping to the Czech language and after the 6-steps re-estimation procedure were tested on two types of databases D1 and D2 (see section: “Input Data”).

	mapped French ph.	re-estimated mapped ph.
phoneme recognition	33% rec 19% acc	65% rec 54% acc
digit recognition	64% rec 63% acc	74% rec 39% acc

Tab. 1 Reached recognition score for both types of models. ”rec” denotes the recognition score, “acc” is the recognition accuracy. In the first column the recognition score with the only mapped phonemes (French phonemes mapped to Czech ones and used for Czech speech recognition) is presented. The recognition results for the bootstrapped Czech phonemes (French phonemes retrained to Czech ones) are given in the second column.

iteration number -	achieved recognition score %
0	33
1	59.97
2	63.31
3	64.83
4	64.91
5	64.81
6	64.54
7	64.49

Tab. 2 Recognition score as the function of BWR steps number. The first line with the 0 iterations is for the mapped models.

- Testing on D1: For the testing of converted phonemes the second half of the training database D1 (other 91

sentences) was used. The results for mapped and re-trained (6-steps phoneme re-estimation procedure used) models can be found in the first row of Tab. 1.

- Testing on D2: The evaluation of converted phonemes was performed on the new database D2 from "real environment" containing 100 digit sequences from 40 speakers. The results can be again found in Tab. 1 (the second row).

Differences in the recognition scores between mapped and re-estimated models (the first vs. second column) can be seen. In both cases the recognition score increases as the result of the used re-estimation procedure. On the other hand, the decreasing of the recognition accuracy (see the second row: 63% for mapped models, 39% for final models) is caused by the inadequate noise model. The original noise model was taken from the database Polyphone and no extra training for this noise model was performed except the re-estimation procedure during bootstrapping process. The re-estimation procedure applied on this noise model is insufficient for the proper noise model matching to the new environment noise containing in the database D2. The inadequate noise model is then the consequence of using two different types of databases. This fact was confirmed e.g. in [18].

3. Conclusion

The described system uses neither more sophisticated models - e.g. diphones, triphones - nor other modifications leading to the recognition score increase. For more information see [11] and [12].

The described approach was used (after further model improvement [14]) for the Czech phonemes evaluation during the solution of word spotting problem and for the construction of phoneme-based speech recognizer working in adverse conditions, e.g. [18], [16]. The described phoneme bootstrap approach was also tested for building speech recognizers for minority languages from the resources such as SpeechDat [17]. In these days, this approach would seem to be not so useful because of the existence of new databases for the Czech language (as e.g. SpeechDatE [19], [20], [21]). But despite of this fact similar approaches have been widely used in systems for speech recognition (see e.g. [16]). Currently very promising is the usage of language independent phonemes for system bootstrap.

Acknowledgement

This work has been (partially) supported by the research program Transdisciplinary Research in Biomedical Engineering No. MSM 210000012 of the Czech University in Prague and by the grant GA 102/02/0124 Voice Technologies for Support of Information Society.

References

[1] NOUZA, J. *InfoCity*. http://itakura.kes.vslib.cz/kes/infocitye.html

[2] HOLADA, M., NOUZA, J. Tools for building, maintaining and evaluating voice operated telephone information system. *13th COST Meeting*. Budapest, 1999 (in the frame of COST 249 – Continuous speech recognition over the telephone line, . http://www.elis.rug.ac.be/ELISgroups/speech/ cost249/index.html).

[3] DOBLER,S. Speech recognition technology for mobile phones. *Ericsson Review*. 2000, no. 3.

[4] ABE, N., WARMUTH, M. K. On the computational complexity of approximating distributions by probabilistic automata. In *Proc. of the 3rd Workshop on Computat. Learning Theory*. 1990, p. 52 – 66.

[5] YOUNG, S. J. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine.* 1996, p. 45 – 57.

[6] RABINER, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE Signal Processing Magazine*. 1989, p 267 – 296.

[7] PICONE, J. Continuous speech recognition using hidden Markov models. *IEEE ASSP Magazine*. 1990, p. 26 – 41.

[8] YOUNG, S. J. *HTK reference manual*. Cambridge University Engineering Department. 1993.

[9] PICONE, J. Signal modelling techniques in speech recognition. *Proceedings of the IEEE*. 1993.

[10] CONSTANTINESCU, A., CHOLLET, G. Swiss polyphone and polyvar: Building databases for speech recognition and speaker verification. *3rd Slovenian-German and 2nd SDRV workshop, Speech and Image Understanding*. Ljubljana (Slovenia), 1996.

[11] SOVKA, P. *Cross language experiment verifying approach converting French flexible vocabulary to Czech vocabulary*. Research report of sabbatical stay. Paris: ENST, 1997.

[12] CONSTANTINESCU, A. et al. Bootstrapping a Czech vocabulary through cross-language use of French phonemes. *Presentation on COST 249 Committee*. Roma (Italy), 1997.

[13] SCHULTZ, T., WAIBEL, A. Experiments on cross-language acoustic modeling. Interactive Systems Laboratories. Carnegie Mellon Univ. (USA), Univ. of Karlsruhe (Germany), *EUROSPEECH 2001*. http://www.is.cs.cmu.edu/papers/speech/EUROSPEECH01/.

[14] VOPIČKA, J. French - Czech cross - language experiment. In *Poster 1998*. Prague: CTU, 1998, p. 46.

[15] VOPIČKA, J., POLLÁK, P., SOVKA, P. UHLÍŘ, J. ASR with noisy speech pre-processing and phoneme model re-estimation. In *Proceedings of Robust Methods for Speech Recognition in Adverse Conditions*. Brussels: COST Office, 1999, p. 151 – 154.

[16] VOPIČKA, J., POLLÁK, P., SOVKA, P. UHLÍŘ, J. Phoneme model based ASR of words in car environment. In *Polish-Hungarian-Czech Workshop on Circuit Theory*, *Signal Processing, and Application.* Prague: CTU, 1999, p. 89 – 92.

[17] JONES, R. J. SpeechDat cymru: A large-scale Welsh telephony database. In *1st International Conference on Language Resources and Evaluation.* Granada (Spain), 1998.

[18] NOVOTNÝ J. *Systematic analysis of triphone-based command recogniser in additive noises*. Internal research report. Prague: FEE CTU, #Z03- 1, 2003 (in Czech)

[19] POLLÁK, P., ČERNOCKÝ, J. *Final recruitement metodology and documentation of speakers typology for the final Czech database*. Research report. Prague: CTU, 2000, ED2.12.2.b. 8 p.

[20] POLLÁK, P., ČERNOCKÝ, J. *Installation of recording device including recording, annotation and documentation on 10 1st speakers – strategy for recruitment of speakers*. Research report. Prague: CTU, 2000, ED.2.12.a. 20 p.

[21] ČERNOCKÝ, J., POLLÁK, P., HANŽL, V. *Czech recordings and annotations on CD's - documentation on the Czech database and database access.* Research rep. Prague: CTU, 2000. ED2.3.2. 37 p.