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

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Maximum F1-Score Discriminative Training for Automatic Mispronunciation Detection in Computer-Assisted Language Learning

Hao Huang[1], *Jianming Wang*[1], *Halidan Abudureyimu*[2]

1. Department of Information Science and Engineering, Xinjiang University, Urumqi, China
2. Department of Electrical Engineering, Xinjiang University, Urumqi, China

{huanghao,jmwang,halidana}@xju.edu.cn

10.21437/Interspeech.2012-248

Abstract

In this paper, we propose and evaluate a novel discriminative training criterion for hidden Markov model (HMM) based automatic mispronunciation detection in computer-assisted pronunciation training. The objective function is formulated as a smooth form of the F_1-score on the annotated non-native speech database. The objective function maximization is achieved by using extended Baum Welch form like HMM updating equations based on the weak-sense auxiliary function method. Simultaneous updating of acoustic model and phone threshold parameters is proposed to ensure objective improvement. Mispronunciation detection experiments have shown the method is effective in increasing the F_1-score, Precision, Recall and detection accuracy on both the training data and evaluation data.
Index Terms: automatic mispronunciation detection, F_1-score, discriminative training, computer-assisted language learning

1. Introduction

Computer assisted language learning that makes use of automatic speech recognition technology has gained a growing interest in the last two decades. Computer Assisted Pronunciation Training (CAPT), which aims at helping the learner by automatically pinpoint erroneous pronunciations, is one of the most popularly deployed applications. Lots of research work has been carried out and various mispronunciation detection techniques have been proposed. While some new paradigm has been explored [1], the HMM based acoustic modeling is the mainstream. Within this framework, the frame-normalized log posterior probability based confidence score is a conventional measurement of correctness.

In HMM based mispronunciation detection, the acoustic models are often trained with maximum likelihood (ML) criterion on native speech data to model the distributions of pronunciation space of the standard speakers. In recent years, discriminative training (DT) has been widely used in speech recognition acoustic model training and has proved to give significant improvement over the traditional ML estimation method. The most common DT methods are minimum classification error (MCE) [2] and maximum mutual information (MMI) [3] or its variant minimum phone error (MPE) [4,5] training. For mispronunciation detection, authors in [6] proposed a discriminative training algorithm to jointly minimize mispronunciation detection errors and diagnosis errors. The acoustic models are refined under the minimum word error criterion. Actually, the above mentioned methods focus on reducing the empirical recognition error (phone error or word error rate) on the training set. In mispronunciation detection task, evaluation measurements of performance can be diversified and entirely different from those used in speech recognition task. The commonly used evaluation criteria might include False Rejection (correct pronunciation detected as incorrect), False Acceptances (errors detected as correct), True Acceptance (correct pronunciation detected as correct) and True Rejection (errors detected as incorrect). Some work use Precision and Recall as the performance measure. These metrics can be effective, however, there has often to be an empirical tradeoff among the multiple objectives.

The F_1-score measure, a synthetic one-dimensional indicator, is nowadays routinely used as an important performance metric to estimate the performance of natural language processing (NLP) or information retrieval (IR) systems. Recently, researchers began to refine system parameters by directly maximizing F_1-score [7,8]. In mispronunciation detection, authors in [9] began to use F_1-score as a performance measurement. On the other hand, the increasing amount of human annotated non-native data makes it reasonable and feasible to directly refine system on large L2 speech corpus. Lots of work has been done such as [1,10]. However, few methods that directly optimize the HMM acoustic models in terms of F_1-score, despite its popularity, have been addressed so far. Inspired by these, we propose a discriminative training algorithm for HMM based automatic mispronunciation which aims at maximizing the empirical F_1-score on the annotated L2 speech data. A smooth version of the F_1-score objective function is proposed, which we denoted as maximum F_1-score criterion (MFC). Extended Baum-Welch (EBW) form like HMM updating functions are derived using the weak-sense auxiliary function method [5]. Mispronunciation detection experiments have shown the effectiveness of the proposed method.

In section 2, the goodness of pronunciation (GOP) based measurement [11] is briefly reviewed. Section 3 discusses the objective function and optimization. Section 4 presents the experiments and the results. Section 5 draws the conclusion.

2. GOP based mispronunciation detection

The task of mispronunciation detection is to verify whether the pronunciation of each phone is correct or not. GOP [11] is the most conventionally used method. In this approach, confusion network which includes canonical phone pronunciations and any possible mispronunciations need to be built. This is normally obtained by force alignment according to the canonical transcription, and then all the possible pronunciation realizations are added. Given a set of acoustic observations of R training utterances: $\mathcal{O}_r, r = 1, \cdots, R$, let $\mathcal{O}_{r,n}$ be the acoustic observations of the nth phonetic segment in utterance r that is composed of N_r segments, and the canonical label of segment (r, n) is denoted as $q_{r,n}$, the GOP of phone segment (r, n) is calculated as:

$$\mathrm{GOP}(\mathcal{O}_{r,n}, q_{r,n}) = \frac{1}{T_{r,n}} \log P(q_{r,n}|\mathcal{O}_{r,n}) \quad (1)$$

$$= \frac{1}{T_{r,n}} \log \left(\frac{p(\mathcal{O}_{r,n}|q_{r,n})P(q_{r,n})}{\sum_{q \in Q(r,n)} p(\mathcal{O}_{r,n}|q)P(q)} \right) \quad (2)$$

where $T_{r,n}$ is the duration (in frames) of the acoustic observation $\mathcal{O}_{n,r}$ in segment (r,n) and $Q(r,n)$ represents all the possible pronunciations of the segment. $q_{r,n}$ is the canonical pronunciation of segment (r,n). Here we assume that all phones share the same prior probability. By using a definition of scaled model probability $g(q,\mathcal{O}) = \kappa \log p(\mathcal{O}|q)$ as the discriminant function, where $0 < \kappa < 1$ is a commonly applied exponential scaling factor in discriminative training to reduce dynamic range of the probabilities, the error detection measurement for segment (r,n) can be written as:

$$d(r,n) = \frac{1}{T_{n,r}} [\log \sum\nolimits_q \exp g(q, \mathcal{O}_{r,n}) - g(q_{r,n}, \mathcal{O}_{r,n})] + \tau \quad (3)$$

$d(r,n) > 0$ is interpreted as the segment $\mathcal{O}(r,n)$ is detected as erroneous and $d(r,n) < 0$ as correct. τ is a global or phone dependent threshold that can be empirically selected or statistically tuned.

3. Maximum F1-score discriminative training

3.1. Objective function

F_1-score is computed from the $Precision$ and $Recall$ of mispronunciations according to the detection results by both the computer and human annotator. F_1-score is computed as:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

where

$$Precision = \frac{N_{WW}}{N_D} \quad Recall = \frac{N_{WW}}{N_W}. \quad (5)$$

N_{WW} is the number of phones marked as error by both the computer and annotator. N_D is the total number of mispronunciations detected by the machine, N_W is the number of mispronunciations judged by the human evaluator. By replacing *Precision* and *Recall* in Eq.(4) with Eq. (5), F_1-score can be reformulated as:

$$F_1 = \frac{2N_{WW}}{N_D + N_W}. \quad (6)$$

In terms of the error detection measurement defined in Eq. (3), N_{WW} and N_D can be expressed as:

$$N_{WW} = \sum\nolimits_{r=1}^{R} \sum\nolimits_{n=1}^{N_r} \mathbb{1}(d(r,n)) \times E(r,n) \quad (7)$$

$$N_D = \sum\nolimits_{r=1}^{R} \sum\nolimits_{n=1}^{N_r} \mathbb{1}(d(r,n)) \quad (8)$$

where $\mathbb{1}(\cdot)$ is the step indicator function defined as:

$$\mathbb{1}(u) = \begin{cases} 1 & if \ u > 0 \\ 0 & if \ u \leq 0 \end{cases} \quad (9)$$

and $E(r,n)$ is the human-annotator result of segment (r,n) and $E(r,n) = 1$ is marked as mispronunciation and 0 otherwise. Note that in Eq. (6), N_W remains a constant, and N_D and N_{WW} are both functions of the HMM parameters λ and we shall maximize F_1-score by optimize λ. Note that F_1-score is not differentiable because of the indicator function. To surpass this problem, we use the sigmoid transfer function to transform the step indicator function into a continuous function similar to MCE training:

$$S(u) = \frac{1}{1 + \exp(-\theta u)} \quad (10)$$

where the constant $\theta > 0$. By replacing the indicator function $\mathbb{1}(\cdot)$ with $S(\cdot)$ in Eq. (7) and (8), a smooth form of F_1-score can be written as follows, which is denoted as *maximum F1-score criterion* (MFC):

$$F_{\mathrm{MFC}} = \frac{2\sum_{r,n} S(d(r,n))E(r,n)}{\sum_{r,n} S(d(r,n)) + N_W} \quad (11)$$

$$= \frac{2N^S_{WW}(\lambda)}{N^S_D(\lambda) + N_W} \quad (12)$$

where $N^S_{WW} = \sum_r^R \sum_n^{N_r} S(d(r,n))E(r,n)$ is the smooth number of errors marked by both the machine and human annotator. $N^S_D = \sum_r^R \sum_n^{N_r} S(d(r,n))$ is the smooth number of errors detected by the machine. From the equation we can see, the objective function can be increased by simultaneously increasing N^S_{WW} and decreasing N^S_D.

3.2. Optimization of the objective function

To optimize the MFC objective function, here we use the weak-sense auxiliary function method which has proved to be successful in MPE training [4,5]. A weak-sense auxiliary function $G(\lambda, \hat{\lambda})$ for the objective function $F(\lambda)$ around current parameter $\hat{\lambda}$ is a smooth function such that [5]:

$$\left.\frac{\partial G(\lambda, \hat{\lambda})}{\partial \lambda}\right|_{\hat{\lambda}} = \left.\frac{\partial F(\lambda)}{\partial \lambda}\right|_{\hat{\lambda}} \quad (13)$$

For the MFC objective function, the weak-sense auxiliary function used can be:

$$G(\lambda, \hat{\lambda}) = \sum_{r=1}^{R} \sum_{n=1}^{N_r} \left.\frac{\partial F_{\mathrm{MFC}}}{\partial \log p(\mathcal{O}_{r,n}|q)}\right|_{\hat{\lambda}} \partial \log p(\mathcal{O}_{r,n}|q) \quad (14)$$

More details about the definition and settings of weak-sense auxiliary function can be seen in [5]. According to the chain rule, the derivative in Eq. (14) can be written as:

$$\frac{\partial F_{\mathrm{MFC}}}{\partial \log p(\mathcal{O}_{r,n}|q)} = \frac{2}{N^S_D(\lambda) + N_W} \frac{\partial N^S_{WW}(\lambda)}{\partial \log p(\mathcal{O}_{r,n}|q)} - \frac{2N^S_{WW}(\lambda)}{(N^S_D(\lambda) + N_W)^2} \frac{\partial N^S_D(\lambda)}{\partial \log p(\mathcal{O}_{r,n}|q)} \quad (15)$$

and the partial derivatives in Eq. (15) are written as:

$$\frac{\partial N^S_{WW}(\lambda)}{\partial \log p(\mathcal{O}_{r,n}|q)} = \frac{1}{T_{r,n}} E(r,n) \times P_1 \quad (16)$$

$$\frac{\partial N^S_D(\lambda)}{\partial \log p(\mathcal{O}_{r,n}|q)} = \frac{1}{T_{r,n}} \times P_1 \quad (17)$$

where

$$P_1 = \frac{\theta e^{-\theta d(r,n)}}{\left(1 + e^{-\theta d(r,n)}\right)^2} \left(\gamma_q(r,n) - \mathbb{1}(q == q_{r,n})\right) \quad (18)$$

And $\gamma_q(r,n)$ is the posterior probability of phone q in segment (r,n). By adding a quantity

$$\gamma_q^{\mathrm{MFC}} = \frac{1}{\kappa} \frac{\partial F_{\mathrm{MFC}}}{\partial \log p(\mathcal{O}_{r,n}|q)} \quad (19)$$

the occupation-data, sum-of-data and sum-of-square-data of the numerator are computed as:

$$\beta_{sm}^{\text{num}} = \sum_{r,n}^{R,N_r} \sum_{t=I(r,n)}^{L(r,n)} \max(0, \gamma_q^{\text{MFC}}) \gamma_{sm}(t) \quad (20)$$

$$\chi_{sm}^{\text{num}} = \sum_{r,n}^{R,N_r} \sum_{t=I(r,n)}^{L(r,n)} \max(0, \gamma_q^{\text{MFC}}) \gamma_{sm}(t) \mathcal{O}(t) \quad (21)$$

$$Y_{sm}^{\text{num}} = \sum_{r,n}^{R,N_r} \sum_{t=I(r,n)}^{L(r,n)} \max(0, \gamma_q^{\text{MFC}}) \gamma_{sm}(t) \mathcal{O}^2(t) \quad (22)$$

where $I(r,n)$ and $L(r,n)$ stand for the initial and last frames of segment (r,n). $\gamma_{sm}(t)$ is the occupation probability of observation $\mathcal{O}_t$ at mixture m of state s and can be computed using forward-backward computation within the segment. The statistics of the denominator β_{sm}^{den},χ_{sm}^{den} and Y_{sm}^{den} can be calculated by replacing $\max(0, \gamma_q^{\text{MFC}})$ with $\max(0, -\gamma_q^{\text{MFC}})$ in Eq. (20-22). In this work we only update the means and covariances of the Gaussians, the iterative updating equations can be written as:

$$\mu_{sm} = \frac{\{\chi_{sm}^{\text{num}} - \chi_{sm}^{\text{den}}\} + D_{sm}\hat{\mu}_{sm}}{\{\beta_{sm}^{\text{num}} - \beta_{sm}^{\text{den}}\} + D_{sm}} \quad (23)$$

$$\sigma_{sm}^2 = \frac{Y_{sm}^{\text{num}} - Y_{sm}^{\text{den}} + D_{sm}(\hat{\sigma}_{sm}^2 + \hat{\mu}_{sm}^2)}{\beta_{sm}^{\text{num}} - \beta_{sm}^{\text{den}} + D_{sm}} - \mu_{sm}^2 \quad (24)$$

where D_{sm} is set to be $D_{sm} = E\beta_{sm}^{\text{den}}$ and the constant E is simply set to be 2 as in MPE training [4,5] in the experiments.

4. Experiments and results

4.1. Database and configurations

The proposed method is evaluated on a Mandarin mispronunciation detection task for Uighur college students who have been learning Putonghua (Mandarin Chinese) in Xinjiang University. The baseline acoustic model is maximum likelihood trained on the '863 project' mandarin speech database. Due to the limited non-native training data, only monophone HMMs are used. The native speech database contains 86 271 utterances read by 160 native Mandarin speakers (80 male and 80 female). The spectral front-end uses 39-dim vector, consisting of 13 MFCCs and their Δ and $\Delta\Delta$ with cepstral mean normalization. The HMM set has 67 phones (28 initials and 37 finals plus silence and short pause), each HMM state is a mixture of 8 Gaussians.

The MFC discriminative training is carried out on a non-native speech corpus. The corpus contains speech data collected from 100 Uighur speakers (50 male and 50 female). Each speaker was asked to read three sets of prompted texts. Each set contains 50 single characters, 25 words and 20 short sentences. The speech has been annotated by well-trained linguists. After cleaning, the database contains of about 25 673 utterances and 18 643 of which are used for MFC training and 7 030 are used for evaluation.

After obtaining the ML trained acoustic models, we then use the acoustic models and canonical phone transcriptions to form a confusion network for each utterance. Each segment of the network includes the canonical phone. The starting/ending time are obtained by using Viterbi alignment. The competitive phone arcs $Q(r,n)$ for the segment are then added. Note that $Q(r,n)$ can be predicted by using handcrafted rules or data driven method or determined by the output of a phone loop recognizer. Here we simply add all the initials into the segment when $q_{r,n}$ is an initial, and otherwise all the finals are added. The probability scaling factor is empirically set to be $\kappa = 0.1$ and the sigmoid constant is $\theta = 10.0$.

Table 1: *F_{MFC} and F_1-score*

	Training		Test	
	F_{MFC}	F_1	F_{MFC}	F_1
Baseline	0.352	0.406	0.344	0.380
MFC only	0.483	0.542	0.367	0.413
MFC ITT-F_{MFC}	0.578	0.704	0.398	0.470
MFC ITT-F_1	0.590	0.718	0.407	0.479

Table 2: *Precision and Recall*

	Training		Test	
	Prec.	*Reca.*	*Prec.*	*Reca.*
Baseline	0.403	0.409	0.352	0.415
MFC ITT-F_1	0.798	0.651	0.518	0.445

Table 3: *Detection error rate (%)*

	Training Set	Testing Set
Baseline	3.75	3.50
MFC ITT-F_1	1.70	2.55

4.2. Results

Table 1 gives a summarization of the results. First we show the baseline result using ML trained acoustic models with threshold τ refined on the L2 training data. The threshold can be a globally set quantity to judge all the phone segments. Because the distribution of each phone's GOP can be variant, we use phone dependent thresholds, which have previously shown to obtain better performance [11]. Ideally, the phone-dependent threshold $\boldsymbol{\tau}$ can be tuned using a grid search process to maximize F_{MFC} on the non-native training data:

$$\boldsymbol{\tau} = \arg\max F_{\text{MFC}} \quad (25)$$

However, grid search can be very time consuming. A simplified technique is adopted: we tune the threshold of each phone q using grid search to find a best F_{MFC} objective function while remaining other phone thresholds fixed. The procedure is repeated until F_{MFC} converges to an optimum. With the MLE trained acoustic model and optimized phone dependent thresholds $\boldsymbol{\tau}$, the baseline F_1-scores on the training and test set are respectively 0.406 and 0.380.

Upon the baseline acoustic models and threshold parameters, 20 iterations of MFC training are conducted which have shown to be sufficient for the objective function to converge. Procedurally, each iteration of MFC training involves the following passes on the training data: (i) Compute the phone arc probabilities of and then the GOP of each aligned segment. (ii) Accumulate the detection statistics such as N_{WW}^S and N_D^S in Eq. (15). (iii) Do forward-backward computations for MFC updating. Table 4 shows the change of F_{MFC} of the training iterations. It can be seen when only the acoustic models are updated, the F_{MFC}, F_1-score on both the training and evaluation set do not improve as much as expected. This is because the distributions of GOP scores on the training data will change after the model parameters λ are updated. Therefore, the thresholds $\boldsymbol{\tau}$ should also be updated correspondingly.

We add an extra threshold searching process after each HMM training iteration, which is denoted as MFC training with *Iterative Threshold Tuning* (MFC ITT-F_{MFC}). It is shown in Table 5 the F_{MFC} objective increases consistently and significantly from 0.352 to 0.578 on the training set. Accordingly, the F_1-score improves from 0.406 to 0.704 on the training data, indicating that the optimization of the objective function F_{MFC} conforms to the optimization of the F_1-score on the training set. For the test set, the F_1-score increases from 0.380 to 0.470, suggesting that optimizing the empirical F_1-score on the training data also leads to F_1-score improvement on the unseen testing

Table 4: *Training iterations of MFC training without ITT*

Iteration	0	2	4	6	8	10	12	14	16	18	20
Train F_{MFC}	0.352	0.395	0.447	0.473	0.464	0.445	**0.483**	0.445	0.450	0.481	0.470
Train F_1-score	0.406	0.436	0.490	0.517	0.520	0.498	**0.542**	0.492	0.505	0.525	0.507
Test F_1-score	0.380	0.367	0.398	0.398	0.412	0.382	**0.413**	0.385	0.398	0.407	0.402

Table 5: *Training iterations of MFC with ITT-F_{MFC}*

Train F_{MFC}	0.352	0.464	0.521	0.521	0.541	0.556	0.560	0.555	0.569	**0.578**	0.573
Train F_1-score	0.406	0.551	0.611	0.618	0.640	0.671	0.672	0.673	0.696	**0.704**	0.703
Test F_1-score	0.380	0.424	0.451	0.452	0.466	0.468	0.469	0.468	0.462	**0.470**	0.468

Table 6: *F_{MBR} of MFC ITT-F_1 training iterations*

Train F_{MBR}	0.552	0.435	0.381	0.376	0.375	0.396	0.388	0.395	0.389	0.382	0.392
Test F_{MBR}	0.564	0.436	0.382	0.371	0.374	0.393	0.383	0.390	0.378	0.378	0.385

data. By comparing the results of MFC ITT-F_{MFC} with those of MFC training without ITT, we can see the importance of simultaneous updating of the HMM and threshold parameters. An alternative of threshold tuning can be accomplished by directly optimizing the F_1-score on the training data:

$$\boldsymbol{\tau} = \arg\max F_1 \tag{26}$$

which is denoted as MFC ITT-F_1. Using MFC ITT-F_1, the F_{MFC}, F_1 on the training set are respectively 0.590 and 0.718 and on the evaluation set are 0.407 and 0.479, slightly but consistently better than those of MFC ITT-F_{MFC}. This might be due to ITT-F_1 is capable of finding a better start point of $\boldsymbol{\tau}$ for the next iteration of MFC training. It is also seen the absolute improvement of F_1-score on the training set (0.312) is much larger than that on the test set (0.100). We attribute this to a sort of overtraining because of the limited L2 training data in this work. We also report Precision and Recall obtained by MFC ITT-F_1 in Table 2. Results suggest that the maximization of the overall metric F_1-score also results in Precision and Recall improvements. Table 3 demonstrates the detection error rate (DER), i.e., the number of segments that have different detection results between the machine and human evaluator normalized by the total number of phone segments N:

$$F_{\mathrm{DER}} = \frac{1}{N}\sum_{r=1}^{R}\sum_{n=1}^{N_r}\mathbb{1}(d(r,n)) \neq E(r,n) \tag{27}$$

Note that F_{DER} can be regarded as an average correct/error classification error rate. It is seen F_{DER} reduces relatively about 54.6% on the training set and 27.1% on the test set, indicating that the improvement of the F_1-score closely correlates with the reduction of the correct/error binary classification errors. Table 6 demonstrates the segmental minimum Bayesian risk (MBR) objective function of the MFC ITT-F_1 training iterations:

$$F_{\mathrm{MBR}} = \frac{1}{N}\sum_{r=1}^{R}\sum_{n=1}^{N_r}\frac{p^{\kappa}(\mathcal{O}_{r,n}|q_{r,n})P^{\kappa}(q_{r,n})}{\sum_{q\in Q(r,n)}p^{\kappa}(\mathcal{O}_{r,n}|q)P^{\kappa}(q)} \tag{28}$$

which can be regarded as a smooth average phone classification accuracy on the L2 data. It is seen though F_1-score increases on both the training set and test set, the smooth expected phone classification accuracy drops significantly. We conjecture that the MBR criterion, which is often applied in speech recognition task, might not be an appropriate objective function for F_1-score optimization in mispronunciation detection.

5. Conclusions

We have evaluated a novel discriminative training criterion which aims at directly maximizing the F_1-score for mispronunciation detection. Objective function maximization is achieved by using EBW form-like HMM updating equations with a simultaneous phone thresholds optimization strategy. Preliminary experiments have shows improvements of F_1-score and other commonly used performance evaluation criteria. The proposed method can be promising with the increasing amount of human annotated L2 speech. Thorough comparisons with previous work such as [6] and with the objective function that minimizes detection error rate, methods to reduce overtraining in DT with low resource L2 data should be remained for the future work.

6. Acknowledgements

This work was supported by the NSFC (60965002, 60865001, 61163026), Scientific Research Program of the Higher Education Institution of Xinjiang (XJEDU2008S15), and Ph.D research fund in Xinjiang university (BS090143).

7. References

[1] Wei S, Hu G P, Hu Y, Wang R H. A new method for mispronunciation detection using Support Vector Machine based on Pronunciation Space Models. *Speech Communication*, 51, 896-905, 2009.

[2] Bahl L R, Brown P F, Souza P, Mercer R. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: Proceedings of *ICASSP*, (1) 49-52, 1986.

[3] Juang B H, Katagiri S. Discriminative learning for minimum error classification. *IEEE Transactions on Signal Processing*, 40(12): 3043-3054, 1992.

[4] Povey D, Woodland P. Minimum phone error and I-smoothing for improved discriminative training. In: Proceedings of *ICASSP* 2002, 105-108.

[5] Povey D. Discriminative Training for Large Vocabulary Speech Recognition, Ph.D. thesis, Cambridge University, 2004.

[6] Qian X, Soong F, Meng H. Discriminative Acoustic Model for Improving Mispronunciation Detection and Diagnosis in Computer-Aided Pronunciation Training (CAPT). In: Proceedings of *Interspeech*, 757-760, 2010.

[7] Fujino A, Isozaki H, Suzuki J. Multi-label Text Categorization with Model Combination based on F1-score Maximization. In: Proceedings of *IJCNLP*, 823-828, 2008.

[8] Dembczynski K, Waegeman W, Cheng W, Hullermeier E. An Exact Algorithm for F-Measure Maximization. In: Proceedings of *NIPS*, 223-230, 2011.

[9] Lo W K, Zhang S, Meng H. Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System. In: Proceedings of *Interspeech*, 765-768, 2010.

[10] Luo Dean, Yang X, Wang L. Improvement of Segmental Mispronunciation Detection with Prior Knowledge Extracted from Large L2 Speech Corpus. In: Proceedings of *Interspeech*, 1593-1596, 2011.

[11] Witt S M, Young S J. Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning. *Speech Communication*, 30 (2-3) 95-108, 2000.