



An open-set detection evaluation methodology applied to language and emotion recognition

David A. van Leeuwen and Khiet P. Truong

TNO Human Factors,
Postbus 23, 3769 ZG Soesterberg, The Netherlands

{david.vanleeuwen, khiet.truong}@tno.nl

Abstract

This paper introduces a detection methodology for recognition technologies in speech for which it is difficult to obtain an abundance of non-target classes. An example is language recognition, where we would like to be able to measure the detection capability of a single target language without confounding with the modeling capability of non-target languages. The evaluation framework is based on a cross validation scheme leaving the non-target class out of the allowed training material for the detector. The framework allows us to use Detection Error Trade-off curves properly. As another application example we apply the evaluation scheme to emotion recognition in order to obtain single-emotion detection performance assessment.

Index Terms: detection methodology, open-set evaluation, language, emotion.

1. Introduction

Speaker Recognition is a speech technology for which the evaluation paradigm is said to be the cleanest of all recognition tasks. Indeed, since the introduction of NIST speaker recognition evaluations in 1996 the task has been specified as a *detection task*, the technology has to decide whether two speech segments have been uttered by the same speaker or not. This beautifully simple statement of the task, together with the potentially very accurate reference truth has made it possible to study various aspects of the speaker recognition task with very high diagnostic accuracy, especially by utilizing the very powerful DET-plot [4] which can integrate an enormous amount of information about many systems or evaluation conditions in a single graph. Because in this case the subject of recognition is a speaker, of whom there is a potential of billions on earth, it is feasible to produce new evaluation databases each year where every speaker, both target and non-target speakers, are ‘new’ and have most likely never been ‘heard’ by the systems under evaluation. This leads to proper separation of target-trial scores from non-target trial scores produced by the system, which are the basis for the DET-plot and for the Detection Cost Function C_{det} , the primary evaluation measure in past NIST-style speaker recognition evaluations¹ [3, 12].

The success of the clean Speaker Recognition Evaluation (SRE) task and the powerful analysis possibilities were transferred to the Language Recognition task, at NIST evaluations held in 1996, 2003 and 2005. However, there is an important difference in the assumption about not having ‘heard’ the target and non-target languages before: in Language Recognition Evaluations (LRE) it assumed that all target languages are

known to the system, and only in some experimental conditions one or more ‘surprise’ languages are used as non-target language. In fact, knowledge of the set of non-target languages is explicitly allowed in NIST-style LREs, which is quite contrary to the SRE where speech of non-target speakers (for a large part available in the ‘training’ part of SRE data) may not be used for trials of a particular target speaker. During the preparations for LRE-2005 Niko Brümmer of Spescom DataVoice showed that the *evaluation priors*, the proportions of trials in the various languages would influence the outcome of C_{det} as defined for LRE at the time [6]. Again, this is different from SRE, where the evaluation priors have no influence to C_{det} , which works with so-called ‘synthetic priors’ characterizing a specific application. This problem of C_{det} was fixed by a fairly complicated new definition [7, 11], but still not all conceptual problems had disappeared. As was shown at the LRE-2005 workshop and discussed at the Odyssey 2006 conference, the interpretation of DET-curves in LRE tasks is unclear.

This paper addresses the difficulty of DET-plots in LRE-detection more in detail in section 2, and then continues with the introduction of a different evaluation methodology in section 3. The methodology does not pretend to be a better alternative to the current NIST LRE evaluation paradigm, but aims at finding a way of being able to make DET plots for LRE. We further show that the methodology can be applied to other areas for which it is difficult to obtain an abundance of ‘unheard’ classes, applying it to emotion recognition in section 4, before stating the conclusions of this paper.

2. Analysis of the problem

As mentioned above, the problem that remained in the LRE was the interpretation of DET curves for LRE. Indeed, almost every site presenting at the LRE-2005 workshop showed DET-plots, including the present author. The conceptual problem lies in the fact that the scores in the target and non-target distributions must be obtained independently. However, because modeling of non-target languages was allowed (and is very beneficial to obtain fewer detection errors), a single test segment would produce 1 target score and $N - 1$ non-target scores that are mutually dependent. Many successful approaches utilize a Gaussian back-end [8, 9], where scores for all potential test languages are converted to *posterior* probabilities, summing to one, which shows their mutual dependence. The conceptual mistake is made, when target scores for different languages (and similarly, non-target scores) are *pooled* to form larger distributions, because information from non-target scores ‘leaks’ to the target scores, and vice versa, due to their dependence.

The problem is most clearly demonstrated for the case of

¹As of 2008, a new, more application-independent evaluation measure called C_{1lr} will be used. For clarity, will use C_{det} in this paper.

$N = 2$ which is applicable to any of the 2-class dialect detection tasks of LRE-2005. Here, each test segment x is tried against both of two possible dialects A and B , and a score is required that expresses the support of the system for the hypothesis that x has dialect A and B , respectively. Now, because this is a two-class discriminative task, posterior-producing systems typically output scores $s_A(x) = p(A|x) = 1 - s_B(x)$, and likelihood-ratio based system will produce something like $s_A(x) = \log p(x|A)/p(x|B) = -s_B(x)$. We will use the latter now for illustration. What happens if we pool all target scores into a single distribution? Each segment a of dialect A will contribute a score in the target distribution, $s_A(a)$ and in the non-target distribution $s_B(a) = -s_A(a)$. Similarly, each segment b of dialect B will give opposite scores in the two distributions. The consequence of this is that the target and non-target distributions are each other's mirror images, as is depicted in Fig. 1. Formally, the score distributions $p(s_A|A) = p(-s_A|B)$ and $p(s_B|B) = p(-s_B|A)$. For the traditional false alarm probability P_{FA}^θ at a threshold θ is calculated as

$$P_{FA}^\theta = \int_{\theta}^{\infty} (p(s_B|A) + p(s_A|B)) ds \quad (1)$$

and the miss rate as

$$P_{miss}^\theta = \int_{-\infty}^{\theta} (p(s_A|A) + p(s_B|B)) ds \quad (2)$$

From these equations and the symmetric relation of the score distributions we get $P_{FA}^\theta = P_{miss}^{-\theta}$. When scores are pooled like this, *symmetric* DET-curves are obtained. False alarms have the roles of misses and vice versa. In [5] most dialect recognition DET curves are symmetric, showing that most systems indeed use a likelihood-ratio or posterior score. The interpretation of these DET curves is problematic. A point of the curve does not represent a single threshold θ , but simultaneously the opposite $-\theta$. Post-evaluation metrics such as C_{det}^{\min} and the Equal Error Rate (EER) have lost their meaning.

Now in this relatively simple case for dialects, the problem due to pooling of score distributions can be resolved, simply by not carrying out the pooling. Then, the axes of the DET plot get a new, discriminative, interpretation: $P(B|A)$ and $P(A|B)$, rather than P_{FA} and P_{miss} . However, in cases with $N > 2$, the matter becomes very complicated. For an inspection of the tradeoff of error probabilities, there are $N - 1$ independent thresholds that may be varied. The concepts EER, DET-plot and C_{det}^{\min} turn out to be far from trivial. The simple question "what would C_{det} be if my thresholds would have been set optimally?" for obtaining C_{det}^{\min} becomes very hard to answer. In [1] options are explored to find such optimum C_{det}^{\min} for the LRE task. In this paper, we take an other approach: we alter the task, to bring it more into conformity with the SRE task.

3. Proposed evaluation methodology

The solution we propose here, is to change the LRE task such that the language detector may use no information from the non-target language in the test². This is accomplished by a cross-validation scheme. The evaluation methodology is as follows. We want to evaluate the performance of a detector for language L_T . We do this, by iterating over non-target languages L_N . For each L_N we require all prior information about L_N to be

²We do not want to suggest that this is a better evaluation methodology than NIST-LRE, in fact, the proposed scheme is not very practical.

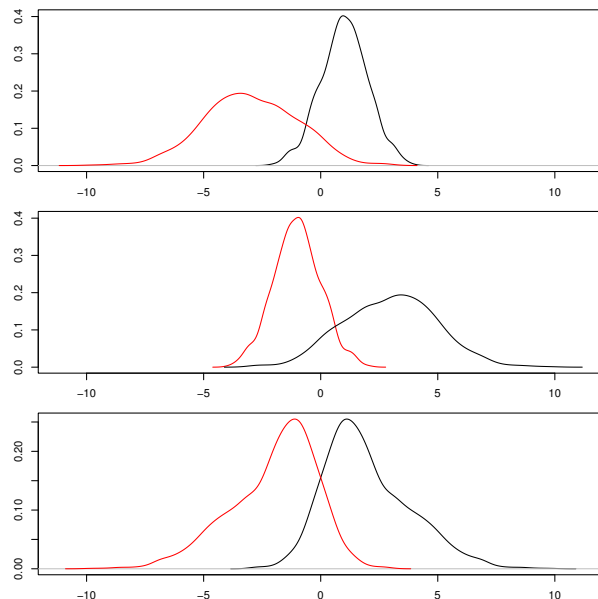


Figure 1: The effect of pooling scores of mutually dependent score distributions, for targets (right) and non-targets (left). Top panel shows scores for targets of class A , middle for targets of class B , and bottom panel the pooled target and non-target scores.

removed from the L_T -detector. Then, the set of available test segments is reduced to only those where the spoken language is either L_T or L_N . A trial is formed by test segment and the question: is this L_T or not, and a score supporting the evidence for L_T is required as well. This process is repeated for all available non-target languages in the evaluation material. All the non-target trial scores are collected, and pooled into a single non-target distribution. The target scores need special attention, because it is likely that there is a large correlation between all cross-validation cases where the test segment x_T is the same, and only the prior information of the detector varies a little bit (excluding knowledge of a different language L_N each time). Two schemes for dealing with these correlated scores for x_T are considered: 1) averaging them, and 2) taking the minimum score, corresponding to the most pessimistic situation. With this scheme, we emulate a one-target language open set detection evaluation. All training information that the detector can use is limited to the target language, and languages other than the non-target language under study.

3.1. Demonstration implementation

We have implemented this cross-validation scheme for N_T using one of our language recognition systems submitted in the NIST LRE-2005. The system consists of 71 GMM/UBM 'score producers,' which are conditioned on the 12 CallFriend languages, 2 genders and 3 channels (for one of the $12 \times 2 \times 3 = 72$ conditions there was no training material found) [11]. A Linear Discriminant Analysis (LDA) back-end [9], trained on lid96d1, lid96e1, lid03e1 and lid05d1 trials³, produces posterior scores for the LRE-2005 languages, given appropriate priors. We implemented 'forgetting' the prior information about L_N in the

³We use the NIST nomenclature *lidyys1*, where *yy* is the LRE year and *s* \in {d, e} denotes development and evaluation data.

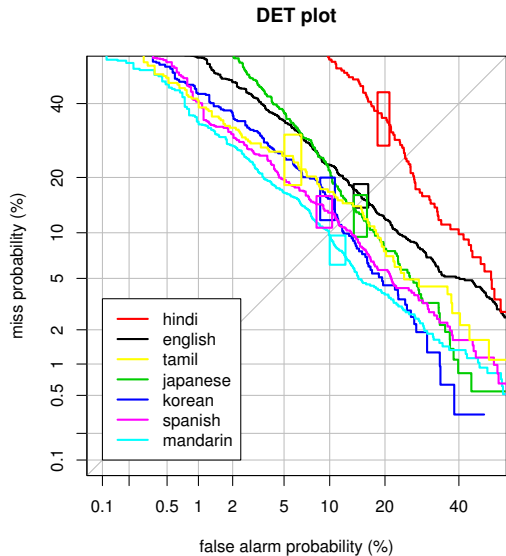


Figure 2: Language detection DET plots for the seven target languages from LRE-2005. The DET curve for English (black) shows the detection capability for English, where the non-English is modeled by all languages except English and the specific non-target language possibly used in the trial.

Table 1: Performance figures of the open-set language detection.

Language	C_{det}	$C_{\text{det}}^{\text{min}}$	EER
Hindi	0.277	0.225	0.246
English	0.157	0.154	0.155
Tamil	0.152	0.131	0.142
Japanese	0.139	0.136	0.137
Korean	0.128	0.110	0.114
Spanish	0.114	0.112	0.113
Mandarin	0.096	0.095	0.099
Mean	0.152	0.138	0.144

following way, by removing from the LDA training: 1) all scores produced by GMMs conditioned on the language L_N (columns in the LDA matrix), 2) all test segments spoken in language L_N (rows of the LDA matrix). Because LDA training can be computed fast, the cross-validation scheme takes very little time compared to a full evaluation. In testing a segment that was either L_T or L_N , GMM scores excluding those conditioned on L_N were selected, and the LDA posterior of L_T was used as the trial score. Decisions were made using prior $p_T = \frac{1}{2}$ for the target language, and $p = 1/2(N-2)$ for the alternative languages from the LDA training set ($N = 12$), and a threshold $\theta = \frac{1}{2}$.

In Figure 2 we show the DET plots obtained from the LRE-2005 evaluation data (all trials, 30 second segments, including ‘surprise’ language German as non-target test segments) using this ‘open set detection’ methodology. We have separated the DET curves for the different target languages. Since the non-target language is always modeled by the set of alternative languages there is no confounding of posterior scores, and the DET curve properly describes the trade-off between P_{FA} and P_{miss} . We can also compute C_{det} and post-evaluation statistics such as the EER and $C_{\text{det}}^{\text{min}}$. In Table 1 we summarize these measures.

The mean of the C_{det} values in Table 1 corresponds to the current definition of C_{det} in NIST LRE, weighting the detec-

tion costs of each target language equally, thereby reducing the influence of evaluation priors. We can compare the mean $C_{\text{det}} = 0.152$ to the performance of the same system *without* the removal of L_N from the system knowledge, $C_{\text{det}} = 0.101$. Clearly, the detection costs can be reduced substantially by explicit modeling of the expected non-target languages.

From the DET plot and the table we can see that the *calibration* of the open-set detectors is pretty accurate, little costs are incurred by suboptimal setting of the threshold.

4. Application to emotion detection

Emotion recognition in speech signals is a relatively new research area, and one of the consequences of this is that there is a limited amount of emotional speech databases available. Often, recordings of actors are made, who express full-blown emotions contrasting a neutral recording. Researchers typically report *classification* experiments on these databases. It is hard to interpret the reported classification performance figures, because they depend, among others, on the type of emotions and the evaluation priors. For a typical application of emotion detection in a call center, where one might want to get statistics of a particular emotion (e.g., anger), a classification experiment of that emotion w.r.t. other very unlikely emotions (happiness, etc.) seems inappropriate. Rather, we would like to have an indication of the performance of a particular target emotion (anger), more or less independent of the other potential emotions in speech.

Because the contrastive emotions of a typical emotion database are in a way ‘arbitrary,’ we would like our emotion detector not to utilize specific knowledge about these non-target emotions. Hence, we can implement the same cross-validation scheme, leaving the non-target emotion out of the training of the target-emotion detector, and representing that specific non-target emotion by the alternative emotions in the database.

We have implemented this cross-validation scheme for the ‘Berlin’ database [2], which contains speech from 10 actors in seven different emotions: neutral (Ne), anger (An), fear (Fe), joy (Jo), sadness (Sa), disgust (Di) and boredom (Bo). Because this is a very small database, and we want to test our emotion detector speaker-independently, a second level of leave-one-out cross-validation is applied by iterating over one test speaker and using the remaining 9 speakers for training of the target and alternative model. We used GMMs for modeling, and a log likelihood ratio as score. We measured C_{det} using a threshold of zero.

In Figure 3 we show the DET curves for these open-set emotion detectors evaluated on the Berlin database, using target scores that were averaged over cross-validation runs. Indicated in this figure are operating points at the EER, which are also tabulated in Table 2. Here we can also compare both discussed methods for dealing with the correlated target scores x_T .

It appears that emotions are difficult to detect if we do not have prior knowledge about the types of potential non-target emotions, except for Sadness which appears to be a very distinct emotion. Fear, in fact, shows the extreme (EER > 50%) where the alternative emotions are very bad representatives of the non-target emotion. In pairwise discrimination experiments performed with this database, it appears that Fear lies right ‘in the middle’ of the other emotions [10].

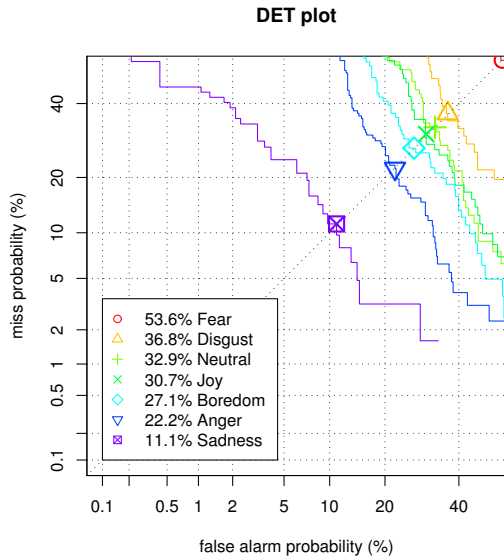


Figure 3: Emotion detection DET plots for seven emotions, emulating open-set detection.

Table 2: Performance figures of the open-set emotion detection. The two rightmost columns represent ‘worst case’ estimates for target scores x_T , see Section 3.

Emotion	Averaging x_T		minimum x_T	
	C_{det}	EER	C_{det}	EER
Fear	0.548	0.536	0.635	0.652
Disgust	0.370	0.368	0.490	0.461
Neutral	0.317	0.329	0.451	0.417
Boredom	0.275	0.271	0.368	0.360
Joy	0.297	0.307	0.353	0.355
Anger	0.214	0.222	0.254	0.252
Sadness	0.117	0.111	0.165	0.129

5. Discussion and Conclusion

We have introduced an evaluation methodology that aims at determining the detection performance of a specific target class, without exploiting knowledge of the non-target class material used in the evaluation data set. This is the natural situation for NIST style speaker recognition evaluation, because all non-target speakers are not known to the system under evaluation. We have applied this methodology to Language Recognition, where we claim to be able to produce a detection performance measure for a particular target language that is less dependent on the alternative languages used in the test database. The methodology simulates a NIST LRE according to [7], but using priors $P_{out-of-set} = 0.5$ and $P_{non-target} = 0$ (NIST nomenclature) instead of 0.2 and $0.3/(N - 1)$ for the open-set condition, respectively. We expect that if a system is optimized using this detection performance measure, it will also perform optimally for NIST-style LRE, which includes known non-targets in the test. Note that this approach does not allow for ‘black box’ evaluation of a system.

We have also applied the methodology to a recognition problem where the choice and definition of the (non-target) classes is somewhat arbitrary, namely that of emotion recognition. Although the detection performances of our system are not particularly good, they are consistent with two-class emo-

tion discrimination experiments [10]. The acoustically ‘most remote’ emotion sadness shows best performance, while the ‘central’ emotion fear is almost impossible to detect acoustically.

6. Acknowledgements

This research is supported by the Dutch BSIK-project MultimediaN <http://www.multimedien.nl>. We would like to thank Niko Brümmer for stimulating discussions.

7. References

- [1] Niko Brümmer and David A. van Leeuwen. On calibration of language recognition scores. In *Proc. Odyssey 2006 Speaker and Language recognition workshop*, San Juan, June 2006.
- [2] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendmeier, and B. Weiss. A database of German emotional speech. In *Proc. Interspeech*, pages 1517–1520, 2005.
- [3] George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds. The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective. *Speech Communication*, 31:225–254, 2000.
- [4] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki. The DET curve in assessment of detection task performance. In *Proc. Eurospeech 1997*, pages 1895–1898, Rhodes, Greece, 1997.
- [5] Alvin F. Martin and Audery N. Le. Current state of language recognition: Nist 2005 evaluation results. In *Proc. Odyssey 2006 Speaker and Language Recognition Workshop*, San Juan, June 2006.
- [6] The 2003 NIST language recognition evaluation plan. <http://www.nist.gov/speech/tests/lang/2003/>, 2003.
- [7] The 2007 NIST language recognition evaluation plan. <http://www.nist.gov/speech/tests/lang/2007/>, 2007.
- [8] Wade Shen, William Campbell, Doug Reynolds Terry Gleason, and Elliot Singer. Experiments with lattice-based PPRLM language identification. In *Proc. Odyssey 2006 Speaker and Language recognition workshop*, San Juan, June 2006.
- [9] Pedro A. Torres-Carrasquillo, Elliot Singer, Mary A. Kohler, Richard J. Greene, Douglas A. Reynolds, and J. R. Deller Jr. Approaches to language identification using gaussian mixture models and shifted delta cepstral features. In *ICSLP*, 2002.
- [10] Khiet P. Truong and David A. van Leeuwen. Visualizing acoustic similarities between emotions in speech: an acoustic map of emotions. In *Proc. Interspeech*, Antwerp, August 2007.
- [11] David A. van Leeuwen and Niko Brümmer. Channel-dependent GMM and multi-class logistic regression models for language recognition. In *Proc. Odyssey 2006 Speaker and Language recognition workshop*, 2006.
- [12] David A. van Leeuwen, Alvin F. Martin, Mark A. Przybocki, and Jos S. Bouten. NIST and TNO-NFI evaluations of automatic speaker recognition. *Computer Speech and Language*, 20:128–158, 2006.