On the Calibration and Fusion of Heterogeneous Spoken Term Detection Systems

Alberto Abad[1], *Luis Javier Rodríguez-Fuentes*[2], *Mikel Penagarikano*[2], *Amparo Varona*[2], *Germán Bordel*[2]

[1]L^2F - Spoken Language Systems Laboratory, INESC-ID Lisboa, Portugal
[2]GTTS, Dept. Electricity and Electronics, University of the Basque Country UPV/EHU, Spain
alberto@l2f.inesc-id.pt, luisjavier.rodriguez@ehu.es

Abstract

The combination of several heterogeneous systems is known to provide remarkable performance improvements in verification and detection tasks. In Spoken Term Detection (STD), two important issues arise: (1) how to define a common set of detected candidates, and (2) how to combine system scores to produce a single score per candidate. In this paper, a discriminative calibration/fusion approach commonly applied in speaker and language recognition is adopted for STD. Under this approach, we first propose several heuristics to hypothesize scores for systems that do not detect a given candidate. In this way, the original problem of several unaligned detection candidates is converted into a verification task. As for other verification tasks, system weights and offsets are then estimated through linear logistic regression. As a result, the combined scores are well calibrated, and the detection threshold is automatically given by application parameters (priors and costs). The proposed method not only offers an elegant solution for the problem of fusion and calibration of multiple detectors, but also provides consistent improvements over a baseline approach based on majority voting, according to experiments on the MediaEval 2012 Spoken Web Search (SWS) task involving 8 heterogeneous systems developed at two different laboratories.

Index Terms: Spoken Term Detection, Majority Voting, Discriminative Calibration and Fusion, MediaEval 2012 SWS.

1. Introduction

Query-by-Example Spoken Term Detection (QE-STD) is a particularly relevant problem for low-resourced languages, which has recently gained interest partially due to the success of the *Spoken Web Search* (SWS) task at the MediaEval evaluation series [1, 2]. In practice, the query-by-example task can be considered as a sort of generalization of the problem of speech search based on text queries, which has been the focus of extensive research activity in the past [3, 4, 5, 6, 7, 8]. In the case of well-resourced languages, a simple straightforward approach to the QE-STD task would first perform speech-to-text conversion of the queries and then apply any of the methods used in text-based speech search. However, when no specific acoustic or lexical knowledge is available for the target languages, like in the SWS task, alternative approaches that do not rely on well-trained acoustic models are needed. In the case of QE-STD, some of the most recent approaches are based on template matching methods, such as different flavours of dynamic time warping (DTW) of posterior derived features [9, 10]. Other systems use acoustic keyword spotting (AKWS) [11, 12], exploiting multilingual acoustic models in several ways. A review of these and other methods can be found in [1, 2].

A common trend in current QE-STD systems is the combination of several (probably weak) detectors, each providing complementary information, which usually leads to improved detection performance. In some cases, the combination is performed at early processing stages, based on the combination of features or DTW matrices (see e.g. [13]). However, when the combination is performed at the final detection stage, the outputs of several heterogeneous systems must be mixed and two important issues arise: (1) how to define a common set of candidate detections (trials), and (2) how to combine system scores to produce a single score per candidate detection. These issues are addressed in [12] by taking only those segments detected by a majority of the systems and averaging the scores of those systems. Under this simple majority voting approach, all the systems are expected to produce scores in the same range, their contributions are equally weighted and the detection threshold is heuristically optimized on a development dataset.

In this paper, a discriminative fusion technique, commonly used in speaker and language verification tasks, is applied for the first time (as far as we know) to the fusion of STD systems. To that end, some heuristic methods are proposed to combine the outputs of several systems, so that a common set of trials can be defined. A majority voting approach such as the one described above is used as baseline and eventually combined with the discriminative fusion to get improved performance. In order to validate the proposed approach, we present results on the SWS challenge included in the MediaEval 2012 evaluation campaign [2]. In this QE-STD task, both the queries and the audio files were extracted from the LWAZI corpus [14], and consist of 8 kHz telephone recordings in four South African languages, containing either read or elicited speech. Separate development and evaluation sets were provided, each containing approximately 1600 audio files and 100 spoken queries. The SWS 2012 official scoring metric based on the Actual/Maximum Term Weighted Value (ATWV/MTWV) [15] is used in this work to assess the systems.

The paper is organized as follows. The problem of fusing several heterogeneous QE-STD systems, along with the baseline and the proposed fusion approaches, are addressed in Section 2. Section 3 describes the QE-STD systems applied in this work, and Section 4 presents and discusses the performance of the baseline and the proposed fusion approaches. Finally, conclusions are given in Section 5.

2. Fusion of heterogeneous STD systems

In speaker and language recognition, the combination of multiple systems at the score detection level using different hetero-

geneous systems is known to provide considerable performance gains [16, 17]. Similarly, STD is expected to benefit from the combination of multiple systems. The approaches typically followed for score fusion in verification tasks assume that a score is output by each system for every possible trial. This is not the case in STD tasks, since each system produces scores for a different set of unaligned candidate detections. Alternatively, some kind of heuristic rule can be used to hypothesize the missing scores for each candidate detection. By doing so, the original detection task is converted into a verification task in which, for each candidate detection (or trial), a score is either output or hypothesized by each system.

2.1. Score normalization

Score normalization methods have proven quite useful in detection tasks such as speaker verification [18]. In QE-STD, system scores may vary depending on the query (for instance, due to different lengths). Moreover, different heterogenous systems may produce scores in different ranges, which in the case of averaging scores to get the fused score may reduce the overall performance, since a single detection threshold is applied. Thus, a kind of normalization is required so that, for any given query and/or subset of systems, the resulting scores are all in the same range. In this work, system scores have been applied a per-query zero-mean and unit-variance normalization (*q-norm*).

2.2. How to hypothesize the missing scores

Detections produced by different QE-STD systems may be unsynchronized/unaligned. Thus, the first issue that must be addressed when fusing several heterogeneous systems is how to align unsynchronized detection scores. First, a list of candidate detections is obtained. Each detection t is associated to a specific query term q_t, an audio file d_t, with initial time i_t, final time f_t, and a vector $\mathbf{s}_t$ containing the scores produced by QE-STD systems. Two candidate detections of a query (produced by two different systems) are considered to be aligned if they (partially) overlap in time. Note that the complete list of detections is equivalent to the list of trials in a verification task, but two important differences can be observed: first, in contrast to a conventional verification task, the whole set of trials is not available beforehand; and second, the score vector $\mathbf{s}_t$ may be sparse, since not all the systems generate scores for every candidate detection t.

In this work, two heuristic criteria have been considered to hypothesize the missing scores: (1) using a *per-query minimum*, i.e. the minimum score produced by the system for that query; or (2) using a *global minimum*, i.e. the minimum score produced by the system for all queries.

2.3. Baseline fusion approach

In [12], three simple heuristic fusion schemes were investigated, among which the Majority Voting (MV) strategy showed the highest and more consistent performance improvements. Under this approach, only candidate detections given by at least half of the systems are kept and the fused score is computed as the mean of the scores for the systems that detected each candidate. In this work, MV is used as the baseline fusion approach.

The MV scheme consists of three decoupled stages: (1) filtering; (2) score hypothesizing; and (3) score fusion. In the filtering stage, some candidate detections are removed. In the score hypothesizing stage, the missing scores are replaced by the mean of the available scores for the considered candidate. Finally, in the fusion stage, system scores are averaged to get the fused score. Note that under this decoupled interpretation of the MV scheme, we could consider alternative score hypothesizing strategies, such as the ones proposed in Section 2.2, or alternative fusion approaches, such as discriminative fusion, which is described in next Section.

2.4. Discriminative calibration and fusion

Under this approach, given N system scores for a candidate detection (trial) t, the fused score is computed as:

$$\hat{s}_t = \beta + \sum_{i=1}^{N} \alpha_i \cdot s_t(i) \tag{1}$$

where the system dependent scaling factors α_i and the offset β are estimated by logistic regression on a development dataset [19, 16]. Logistic regression estimations lead to improved discriminative and well-calibrated scores that approximate the log-likelihood ratio:

$$\hat{s}_t \approx \log \frac{P(\hat{s}_t|H_{\text{target}})}{P(\hat{s}_t|H_{\text{non-target}})} \tag{2}$$

where H_{target} and $H_{\text{non-target}}$ denote target and non-target hypotheses, respectively. Well-calibrated scores allow the use of theoretically determined decision thresholds that only depend on the prior and costs (the operating point) of the evaluation cost function.

For example, in past NIST Speaker Recognition Evaluations, the so called Detection Cost Function was used:

$$C_{\text{Det}} = C_{\text{miss}} \cdot P_{\text{miss}} \cdot P_{\text{tar}} + C_{\text{fa}} \cdot P_{\text{fa}} \cdot (1 - P_{\text{tar}}) \tag{3}$$

where C_{miss} and C_{fa} are the costs of detection errors and P_{tar} is the *a priori* probability of the target speaker. For such cost function, and given well-calibrated scores, the theoretical minimum expected cost Bayes threshold would be:

$$\theta_{\text{Bayes}} = \log \frac{C_{\text{fa}} \cdot (1 - P_{\text{tar}})}{C_{\text{miss}} \cdot P_{\text{tar}}} \tag{4}$$

In Mediaeval 2012 SWS, the Term-Weighted Value was used as the evaluation metric:

$$TWV = 1 - \underset{term}{\text{average}} \{P_{\text{miss}}(term) + \beta P_{\text{fa}}(term)\} \tag{5}$$

Note that maximizing TWV is equivalent to minimizing $1 - TWV$. Assuming that all the terms are equally likely, an equivalent cost function would be:

$$C = P_{\text{miss}} + \beta P_{\text{fa}} \tag{6}$$

which can be seen as a particular case of the Detection Cost Function (for $C_{\text{miss}} = 2$, $C_{\text{fa}} = 2\beta$ and $P_{\text{tar}} = 0.5$). Thus, the theoretical minimum expected cost Bayes threshold would be:

$$\theta_{\text{Bayes}} = \log \beta \tag{7}$$

In a first attempt to apply discriminative fusion to QE-STD systems, well-calibrated scores were not obtained (the theoretical threshold fell far from the optimum). After a first analysis, we noticed that logistic regression estimations must rely on the full set of trials, but QE-STD systems generate just a few of them. Therefore, the full set of trials was generated, by hypothesizing the missing scores as stated in Section 2.2. Figure 1 shows the TWV curve along with the theoretical Bayes threshold on the evaluation set of the SWS task, for the fusion of 8 systems under a combination of MV and discriminative fusion (see Section 4.2 for details). Note that the theoretical Bayes threshold matches almost perfectly the optimal threshold obtained empirically.

In this work, discriminative calibration and fusion have been estimated and applied by means of the Bosaris toolkit [20].

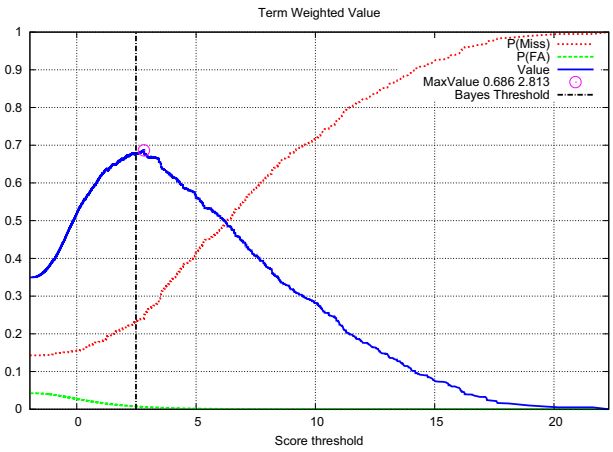

Figure 1: TWV, P_{miss} and P_{fa} curves, maximum TWV and theoretical Bayes threshold for a combination of MV and discriminative fusion of 8 systems on the evaluation set of the SWS task.

3. QE-STD systems

3.1. L^{2}F systems

L^{2}F systems are based on hybrid connectionist methods [21] for both query tokenization and search. Four individual sub-systems have been developed exploiting four different language-dependent acoustic models trained for European Portuguese (PT), Brazilian Portuguese (BR), European Spanish (ES) and American English (EN). The acoustic models from each system are in fact multi-layer perceptron (MLP) networks that are part of L^{2}F in-house hybrid connectionist ASR system named AUDIMUS [22].

3.1.1. The baseline connectionist ASR system

Speech recognizers combine four MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), PLP with log-RelAtive SpecTrAl speech processing features (PLP-RASTA, 13 static + first derivative), Modulation SpectroGram features (MSG, 28 static) and Advanced Font-End from ETSI features (ETSI, 13 static + first and second derivatives). The language-dependent MLP networks were trained using different amounts of annotated data. Each MLP network is characterized by the input frame context size (13 for PLP, PLP-RASTA and ETSI; 15 for MSG), the number of units of the two hidden layers (500), and the size of the output layer. In this case, only monophone units are modeled (EN: 41, PT: 39, BR: 40 and ES: 30). The decoder is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition [23, 24].

3.1.2. Spoken Query tokenization and search

The phonetic transcription of each spoken query is obtained for every sub-system using a phone-loop grammar with phoneme minimum duration of three frames. Simple *1-best* phoneme chain output has been used. Then, search is carried out with a sliding window of 5 seconds (2.5 seconds time shift) using an equally-likely 1-gram language model formed by the target query and a competing speech background model. On the one hand, keyword/query models are described by the sequence of phonetic units obtained in the tokenization. On the other hand, the likelihood of a background speech unit representing "general speech" is estimated based on the other phonetic classes [25, 26]. The output score for each candidate detection is computed as the average of the phonetic log-likelihood ratios that form the detected query term.

3.2. GTTS systems

3.2.1. Feature extraction

GTTS systems used a frame-level sequence of phone log-likelihoods to represent both the query and the audio document. The open software Temporal Patterns Neural Network (TRAPs/NN) phone decoders, developed by the Brno University of Technology (BUT) for Czech (CZ), Hungarian (HU) and Russian (RU) [27] were applied to get frame-level phone posterior probabilities at a rate of 100 frames per second. Feature vectors consisted of 43, 59 and 50 log-likelihoods for the systems based on the Czech, Hungarian and Russian decoders, respectively. A fourth system (called B3) was also developed by concatenating log-likelihood features for the three BUT decoders.

3.2.2. Spoken Query search

Based on the above described representation, multiple occurrences of the spoken query inside an audio document were detected just by defining a cosine distance measure between two feature vectors and recursively applying a Dynamic Time Warping (DTW) approach which minimized the length-normalized accumulated distance, based on a distance matrix with query-normalized distances (each element ranging from 0 to 1). All the frames of the audio segment were explored as initial points of a match, and the distance was accumulated until the optimal aligment reached the last frame of the spoken query. A number of non-matching frames were skipped before and after the best match. The output score for a match was computed as 1 minus the length-normalized accumulated distance. Length and distance normalizations, along with speech/non-speech detection (not described here for lack of space), were key for detection performance.

4. Results

Table 1 shows the performance achieved by the 8 single systems considered in this work, 4 by GTTS (B3, CZ, HU and RU) and 4 by L2F (BR, EN, ES and PT), on the development (dev queries and dev audio files) and evaluation (eval queries and eval audio files) datasets of the MediaEval 2012 SWS task. Query normalization has been applied to all sub-systems and it is also applied in the remaining experiments, since it consistently led to improved performance in all cases. Regarding these results, the MTWV on the evaluation dataset ranged from 0.296 for the EN system to 0.505 for the B3 system, whereas the ATWV ranged from 0.295 to 0.495 for the same systems, showing that the optimal threshold found heuristically on the development dataset matched almost perfectly the evaluation dataset.

4.1. Majority Voting vs. Discriminative Fusion

The baseline fusion approach (*MV*), based on majority voting and score averaging, was applied to three sets of systems: GTTS, L2F and GTTS+L2F. As shown in Table 2, performance improved in all cases, specially when the fusion involved heterogeneous systems (by *heterogeneous* we mean systems applying different methodologies to search for spoken queries). Taking the best single system of each set as reference, ATWV improvements on the evaluation dataset were of around 4.9% for GTTS, 8.6% for L2F and 26.0% for GTTS+L2F. Once again,

Table 1: MTWV/ATWV performance for single STD systems. ATWV is shown for the optimal heuristic threshold in dev-dev.

System	dev-dev		eval-eval	
	MTWV	ATWV	MTWV	ATWV
B3	0.478	0.478	0.505	0.495
BR	0.405	0.405	0.409	0.408
CZ	0.357	0.357	0.393	0.374
EN	0.275	0.275	0.296	0.295
ES	0.348	0.348	0.395	0.391
HU	0.323	0.323	0.352	0.329
PT	0.439	0.439	0.471	0.469
RU	0.403	0.403	0.390	0.389

Table 2: MTWV/ATWV performance for the fusion of three sets of STD systems under the baseline MV approach. ATWV is shown for the optimal heuristic threshold in dev-dev.

System	dev-dev		eval-eval	
	MTWV	ATWV	MTWV	ATWV
GTTS	0.493	0.493	0.526	0.520
L2F	0.535	0.535	0.521	0.510
GTTS-L2F	0.598	0.598	0.628	0.624

the optimal threshold on development proved to be also adequate for evaluation.

The proposed fusion approach (*DF*) was applied to the same sets of systems (GTTS, L2F and GTTS+L2F), using two different methods to hypothesize the missing scores (*QMin* and *GMin*, corresponding to the per-query minimum score and the global minimum score, respectively). As shown in Table 3, DF outperformed MV-based fusions and provided well-calibrated scores in most cases. Note that in this case the Bayes optimal threshold (automatically derived from application costs and priors) was applied to compute ATWV on the evaluation dataset. When using *QMin*, ATWV improvements over the MV-based fusion were of around 2.6% for GTTS, 4.1% for L2F and 4.3% for GTTs+L2F. When using *GMin*, the L2F system was poorly calibrated (an open issue we are currently investigating). However, the MTWV was remarkably high for GTTS+L2F, yielding a 7% improvement over the MV-based fusion and revealing the potential of *GMin* provided that calibration issues are solved.

4.2. Taking the best of both: Integrated approach

An integrated fusion approach (*MV+DF*) was applied to the GTTS+L2F set of systems, using the two hypothesizing methods described above. The integrated approach consists of a first MV filtering to remove some candidate detections followed by discriminative fusion. MTWV and ATWV performance on the evaluation dataset of MediaEval 2012 SWS is shown in Figure 2, and compared to MV-based fusion for different values of m: the number of systems required to detect a candidate. Note that the baseline MV fusion corresponds to $m = 4$ and the discriminative fusion as applied in Section 4.1 corresponds to $m = 1$. Once again, discriminative fusion outperforms MV in all cases, but the most remarkable result is that MV filtering helps to improve the performance of discriminative fusion. The best figure is attained for DF *GMin* with $m = 3$ (ATWV=0.678), providing a 4.1% additional improvement over the best discriminative fusion with no filtering of candidate detections (ATWV=0.651, for $m = 1$ and *QMin*).

Table 3: MTWV/ATWV performance for the fusion of three sets of STD systems under the proposed discriminative fusion approach, using two different methods to hypothesize the missing scores. ATWV is shown for the Bayes optimal threshold.

Fusion mode	System	dev-dev		eval-eval	
		MTWV	ATWV	MTWV	ATWV
Qmin	GTTS	0.505	0.500	0.538	0.533
	L2F	0.540	0.531	0.549	0.531
	GTTS-L2F	0.644	0.640	0.654	0.651
Gmin	GTTS	0.505	0.500	0.536	0.533
	L2F	0.539	0.533	0.518	0.439
	GTTS-L2F	0.659	0.646	0.672	0.624

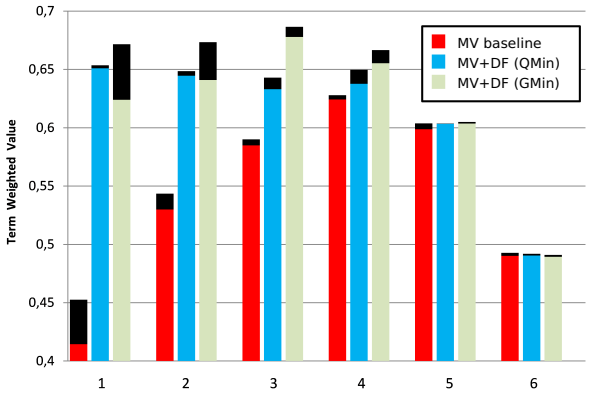

Figure 2: ATWV (color) and MTWV (black) performance for the fusion of GTTS+L2F systems using MV and two variations of the integrated fusion approach, for different number of systems required to detect a candidate.

Note that Bayes thresholds were applied for the DF-based methods (blue and light-green bars) resulting in particularly good calibration for the *Qmin* case (blue bar). Finally, DF was less sensitive to m, providing significantly better results than the baseline MV method for lower m values.

5. Conclusions

In this paper, a discriminative fusion approach commonly applied in verification tasks has been investigated and introduced for the first time to Spoken Term Detection. Results on the MediaEval 2012 SWS task confirm that the proposed approach outperforms a baseline Majority Voting fusion. Besides, it provides calibrated scores for which a theoretical optimum Bayes threshold can be used for making hard decisions. An integrated approach that comprised MV filtering, hypothesizing of missing scores and discriminative fusion yielded even better results. Current work involves solving calibration issues observed for some configurations and finding new ways of hypothesizing the missing scores.

6. Acknowledgements

This work has been partially funded by the DIRHA European project (FP7-ICT-2011-7-288121), the Portuguese Foundation for Science and Technology (FCT) through the project PEst-OE/EEI/LA0021/2013, the University of the Basque Country under grant GIU10/18 and project US11/06, and the Government of the Basque Country under program SAIOTEK (project S-PE12UN055).

7. References

[1] F. Metze, N. Rajput, X. Anguera, M. Davel, G. Gravier, C. van Heerden, G. V. Mantena, A. Muscariello, K. Prahallad, I. Szöke, and J. Tejedor, "The Spoken Web Search Task At MediaEval 2011," in *Proc. ICASSP*, 2012.

[2] F. Metze, X. Anguera, E. Barnard, M. Davel, and G. Gravier, "The Spoken Web Search Task At MediaEval 2012," in *Submitted to Proc. ICASSP*, 2013.

[3] M. Weintraub, "LVCSR log-likelihood ratio scoring for keyword spotting," in *in Proc. ICASSP*, 1995.

[4] I. Szöke, P. Schwarz, P. Matějka, and M. Karafiát, "Comparison of keyword spotting approaches for informal continuous speech," in *In Proc. Eurospeech*, 2005.

[5] J. Garofolo, C. Auzanne, and E. Voorhees, "The TREC Spoken Document Retrieval Track: A Success Story," in *in Text Retrieval Conference (TREC) 8*, 2000.

[6] K. Ng and V. Zue, "Subword-based approaches for spoken document retrieval," *Speech Communication*, vol. 32, no. 3, pp. 157–186, 2000.

[7] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proc. SIGIR*, 2007.

[8] D. Miller, M. Kleber, C.-L. Kao, O. Kimball, T. Colthurst, S. Lowe, R. M.Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Proc. Interspeech*, 2007.

[9] X. Anguera, "Telefonica system for the spoken web search task at MediaEval 2011," in *Proc. MediaEval Workshop*, 2011.

[10] A. Muscariello, G. Gravier, and F. Bimbot, "A zero-resource system for audio-only spoken term detection using a combination of pattern matching techniques," in *Proc. Interspeech*, 2011.

[11] I. Szöke, J. Tejedor, M. Fapso, and J. Colás, "BUT-HCTLab approaches for spoken web search," in *Proc. MediaEval Workshop*, 2011.

[12] A. Abad and R. F. Astudillo, "The L^2F Spoken Web Search system for MediaEval 2012," in *Proc. MediaEval 2012 Workshop*, 2012.

[13] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Using parallel tokenizers with DTW matrix combination for low-resource Spoken Term Detection," in *Proc. ICASSP*, 2013.

[14] E. Barnard, M. Davel, and C. van Heerden, "ASR corpus design for resource-scarce languages," in *Interspeech*, 2009.

[15] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, "Results of the 2006 Spoken Term Detection Evaluation," in *Proc. SIGIR*, 2007.

[16] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[17] L. Rodríguez, M. Peñagarikano, A. Varona, M. Díez, G. Bordel, D. Martínez, J. Villalba, A. Miguel, A. Ortega, A. Lleida, E.and Abad, O. Koller, I. Trancoso, P. Lopez-Otero, L. Fernández, C. García-Mateo, R. Saeidi, M. Soufifar, T. Kinnunen, T. Svendsen, and P. Fränti, "Multi-site heterogeneous system fusions for the Albayzin 2010 Language Recognition Evaluation," in *Proc. ASRU*, 2011.

[18] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.

[19] N. Brummer and D. Van Leeuwen, "On calibration of language recognition scores," in *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.

[20] N. Brummer and E. de Villiers, "The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing," Tech. Rep., 2011. [Online]. Available: https://sites.google.com/site/nikobrummer

[21] N. Morgan and H. Bourlad, "An introduction to hybrid HMM/connectionist continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 25–42, 1995.

[22] H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso, and J. Neto, "The L^2F Broadcast News Speech Recognition System," in *Proc. Fala*, 2010.

[23] M. Mohri, F. Pereira, and M. Riley, "Weighted Finite-State Transducers in Speech Recognition," *Computer Speech and Language*, vol. 16, pp. 69–88, 2002.

[24] D. Caseiro and I. Trancoso, "A Specialized On-The-Fly Algorithm for Lexicon and Language Model Composition," *IEEE Transactions on Audio, Speech and Lang. Proc.*, vol. 14, no. 4, Jul. 2005.

[25] J. Pinto, A. Lovitt, and H. Hermansky, "Exploiting phoneme similarities in hybrid HMM-ANN keyword spotting," in *Proc. Interspeech*, 2007, pp. 1817–1820.

[26] A. Abad, A. Pompili, A. Costa, and I. Trancoso, "Automatic word naming recognition for treatment and assessment of aphasia," in *Proc. Interspeech*. ISCA, September 2012.

[27] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/, Brno, Czech Republic, 2008.