ATVS-UAM NIST LRE 2009 SYSTEM DESCRIPTION

Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Javier Franco-Pedroso, Daniel Ramos, Doroteo T. Toledano, and Joaquin Gonzalez-Rodriguez

ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain
{ignacio.lopez, javier.gonzalez, javier.franco, daniel.ramos, doroteo.torre, joaquin.gonzalez} @uam.es

Submission Overview

ATVS submission to LRE'09 consists of four different combinations of acoustic and phonotactic subsystems. The two ATVS acoustic subsystems are based in session variability compensated sufficient statistics, the first system was built according to the FA-GMM linear scoring framework and the second one is a SVM whose inputs are model supervectors adapted from the compensated sufficient statistics. The phonotactic components are PhoneSVM composed of seven ATVS and three BUT tokenizers. Dual models are obtained in the front end for VOA (22 models, indian-english not trained because of data scarcity) and CTS (14 models) data, while all submissions use an anchor model back-end (23 VOA+CTS models, indian-english learned from other 22 model scores). Front-end scores are channel dependent (22 VOA/14 CTS) t-normalized while back-end scores are channel-independent (23 VOA+CTS) t-normalized and duration-dependent (30s-10s-3s) calibrated. Output scores are submitted in the form of log-likelihood ratio (logLR) scores in an application independent way. Those logLRs have been developed in order to minimize a one-vs.-all Cllr per target language, through 23 language-dependent calibration processes trained considering 23 independent one-vs.-all duration-dependent detection problems per system. In order to train the calibration for development, a cross-validation scheme has been used, which allowed the efficient use of the available development scores. Closed-set detection thresholds have been set to the one-vs.-all Bayes thresholds in all cases (trained with the available closed-set data), and the same logLR sets are submitted to the closed- and open-set conditions. Language-pair llrs and decisions are directly submitted from the closed-set llr scores. ATVS and TNO have shared scores for their respective primary systems: ATVS dot-scoring scores have been provided to TNO, and we include in our ATVS1 primary system, together with all ATVS acoustic and phonotactic subsystems, TNO 3 dot-scoring and 3 Supervector-GMM (obtained with 3 different UBMs) systems. Contrastive systems are ATVS2 (all and only ATVS systems), ATVS3 (ATVS dot scoring alone) and ATVS4 (ATVS 10 PhoneSVMs). A closed-set development dataset, known as ATVS-Dev09, composed of portions or all of LRE'05, Callfriend, LRE'07 and VOA data (different portions and/or selection criteria for train and test) has been used to test the submitted systems in the 23 languages of LRE'09. Development results using the ATVS-Dev09 set for the closed-set 30s condition for the ATVS1/2/3/4 submitted systems yield Cavg (%) values of 1.8, 2.5, 3.7 and 4.2 respectively.

1. System ATVS1 (Primary System)

1.1. System description

Our systems are based on different anchor model back-end combinations of several subsystems. First we will describe the individual sub-systems in sections 1.1.1 to 1.1.4, and then we will describe the fusion of the individual subsystems in section 1.1.5.

1.1.1. DS-CS: FA-GMM linear scoring system

ATVS DS-CS (DotScoring with Compensated Statistics) GMM-FA linear scoring system is based on the work carried out by Niko Brummer and Albert Strasheim for the past NIST speaker recognition evaluation [SRE08]. Details of these systems can be founded here [Strasheim 08]. In this work a complete acoustic system based on generative modelling GMM-FA framework is introduced, adding a new scoring approach based on a linear aproximation to log-likelihood ratios. System shows a great performance in both computational and detection costs.

Parameterization is shared among acoustic systems, consisting in 7 MFCCC with CMN-Rasta-Warping concatenated to 7-1-3-7 SDC-MFCCs. Given a UBM, sufficient stats are extracted for every utterance (train and test); then, first stats are session variability compensated following the FA recipe and models are generated from the compensated stats. Finally scores are obtained via dot product between test first compensated stats and model supervector.

Session variability subspace (U) was trained via EM algorithm after a PCA initialization based on the works [Kenny 05][Vogt 08], where only top-50 eigenchannels were taken into account.

Two different GMM-FA linear scoring systems were developed according to the two different type of data presented in the evaluation. In that sense two UBMs and U matrices are trained from telephone and broadcast data respectively. We found this approach to outperform the approach where mixed data (CTS, broadcast) is processed to train a unique session variability subspace.

UBM-CTS (M=1024) is obtained from CallFriend, LRE’05 and TrainLRE’07 data (165 minutes/language x 14 languages =38,5 hours) while UBM-VOA (M=1024) is obtained from 22 (all VOA languages except Indian English) VOA2+VOA3 data (85 minutes/language x 22 languages = 31,2 hours). U-CTS and U-VOA are obtained from 600 150-second files per language (U-CTS: 14x150sx600 = 350 hours; U-VOA: 22x150sx600 = 550 hours).

1.1.2. SV-CS: SVM chanel compensated supervector

ATVS supervector approach is also based on the stats computed in 1.1.1. In this case every compensated sufficient stats is adapted from the UBM model (trained with the same data as 1.1.1 but M=512). Therefore we obtain a single adapted stat per utterance that summarises its information. Difference between the standard supervector, and stats-based supervector is that in the latter case we replace the vector of means of the adapted GMM by the utterance adapted stats.

1.1.3. PhX: Phone-SVMs

Each of the seven different ATVS Phone-SVM subsystems is based on the following steps. First a voice activity detector segments the test utterance into speech and non-speech segments. The speech segments are recognized with one open-loop phonetic decoder. The best decoding is used to estimate count-based 1-grams, 2-grams and 3-grams, pruned with a probability threshold, resulting in about 40.000 ngrams per recognizer. All these parameters are reshaped as a single vector that is taken as the input of an SVM that classifies the test segment as corresponding (or not) to one language.

The process described above is repeated for the seven different open-loop phonetic recognizers used. In particular these subsystems use six phonetic decoders trained on SpeechDat-like corpora, each of which contain over 10 hours of training material covering hundreds of different speakers. The languages of these phonetic decoders and the corresponding corpora used are English (with the corpus with ELDA catalogue number S0011), German (S0051), French (S0185), Arabic (S0183 + S0184), Basque (S0152) and Russian (S0099). We have also included a 7th phonetic decoder in Spanish trained on Albayzin [Moreno 93] downsampled to 8 kHz, which contains about 4 hours of speech for training. All these decoders are based on Hidden Markov Models (HMMs) trained using HTK and used for decoding with SPHINX. The phonetic HMMs are three-state left-to-right models with no skips, being the output pdf of each state modeled as a weighted mixture of Gaussians.

The acoustic processing is based on 13 Mel Frequency Cepstral Coefficients (MFCCs) (including C0) and velocities and accelerations for a total of 39 components, computing a feature vector each 10ms and performing Cepstral Mean Normalization (CMN).

For each test utterance, the systems make n-grams with the 1-best solution produced by the phonetic decoders. Support Vector Machines (SVMs) take the n-grams as input vectors [Campbell 06].

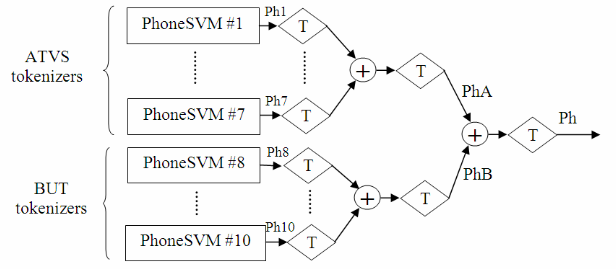

Fig. 1: *Hierarchical combination of phonotactic systems. T stands for t-norm, performed in a channel dependent way (VOA/CTS) in front-end systems.*

Additionally, three speech recognizers (Hungarian, Czech and Russian) from BUT (Speech@FIT, Speech Processing Group at Faculty of Information Technology, Brno University of Technology - FIT BUT, Czech Republic) have been used as additional high-quality tokenizers. The PhoneSVM systems are built then in the same way as with ATVS tokenizers. PhoneSVMs are combined in different ways to obtain different Front-end systems, as shown in fig. 1. Each PhX system consists of 22 VOA and 14 CTS models trained separately. Channel dependent t-norm is the last stage of those phonotactic front-ends.

1.1.4. TNO Acoustic Systems

TNO has contributed to our primary submission with scores from three versions (from 3 different UBMs) of two systems, namely a FA-GMM linear scoring system (TNO-DS-UBM1/2/3) and a Supervector-GMM (TNO-SV-UBM1/2/3). TNO system details are to be found in their system description submitted to this eval.

1.1.5. Primary system: ATVS1

Our back-end strategy for this eval is based on the use of anchor models [Lopez 08], where high-dimensionality input vectors are classified in a single SVM per target model (23) both for VOA and CTS data. Recently, the anchor models approach has been successfully used for speaker verification and language identification too [Collet 05][Noorl 06]. By using anchor models, each utterance is mapped into a model space where the relative behaviour of the speech utterance with respect to other models can be learned. The mapping function consists of testing every single utterance over a cohort of reference models, known as anchor models. The feature vector is the concatenation of all the scores. Input vectors to our primary back-end have dimension 438 (36 ATVS models -14CTS+22VOA- x 6 component systems + 37 TNO models -14CTS+23VOA- x 6). Back-end t-norm is channel-independent (VOA+CTS), while calibration is duration-dependent. Anchor model training is 90/10 bootstrapped while calibration training is bootstrapped with 80/20.

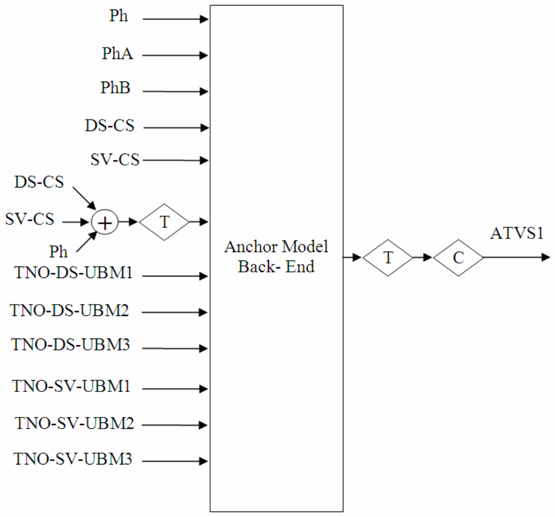

Fig. 2: *ATVS1 Primary system. T stands for t-norm, performed in a channel independent way (VOA+CTS) in back-end systems. Calibration (C) is duration dependent (30s-10s-3s).*

1.2. Training data used

A closed-set development dataset, known as ATVS-Dev09, composed of portions or all of LRE'05, Callfriend, LRE'07 and VOA data (different portions and/or selection criteria for train and test and for each language) has been used to test the submitted systems in the 23 languages of LRE'09.

The training material for the CTS language models consisted of the Callfriend database, the full-conversations of NIST LRE 2005 and development data of NIST LRE 2007. For Russian data we used also RuSTeN (LDC 2006S34 ISBN 1-58563-388-7). VOA models are obtained from speech segments (min. length 30 s.) extracted from VOA2 and VOA3 long files (except manually labeled files, used for testing) using telephone labels distributed by BUT.

Training of the phonetic models used in the ATVS PhoneSVM systems is described in section 1.1.3.

1.3. Processing speed

See Annex 1.

2. Contrastive Systems: ATVS2/3/4

2.1. System description

Contrastive systems make use of the same ATVS individual sub-systems described in Section 1, combined as follows.

2.1.1. ATVS2: phonotactic + acoustic

ATVS2 is strongly similar to ATVS1 but avoiding any TNO system and including individual PhoneSVMs 1 to 10, as shown in figure 3. Input vectors to our ATVS2 back-end have dimension 576 (36 models x 16 component systems).

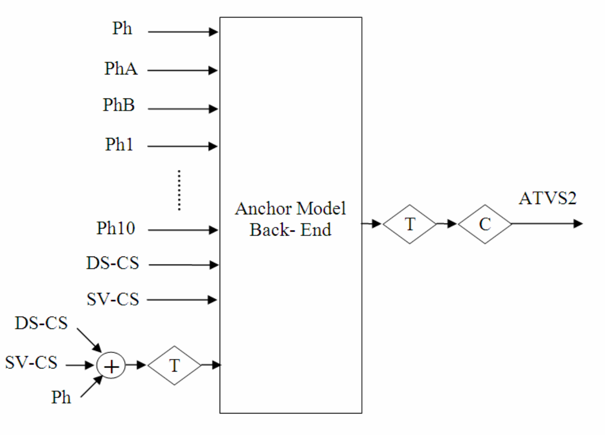

Fig. 3: *ATVS2: ATVS-only PhoneSVMs plus acoustic systems.*

2.1.2. ATVS3: FA-GMM linear scoring system

ATVS3 is a fast and simple acoustic system, as shown in figure 4. Input vectors to ATVS3 back-end have dimension 36 (36 models x 1 component systems).

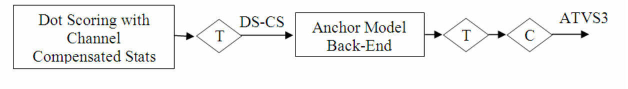

Fig. 4: *ATVS3 contrastive acoustic system.*

2.1.3. ATVS3: phonotactic system

ATVS3 is a combination of all ATVS PhoneSVMs, as shown in figure 4. Input vectors to ATVS4 back-end have dimension 468 (36 models x 13 component systems).

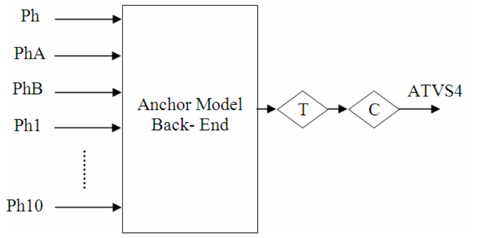

Fig. 5: *ATVS4 contrastive phonotactic system.*

2.2. Training data used

The training data used for those systems is exactly the same as described for our primary system (see section 1.2).

2.3. Processing speed

See Annex 1.

3. Development Results

A closed-set development dataset, known as ATVS-Dev09, composed of portions or all of LRE'05, Callfriend, LRE'07 and VOA data (different portions and/or selection criteria for train and test and for each language) has been used to test the submitted systems in the 23 languages of LRE'09.

The training material (ATVS-DevTrain09) for the CTS language models consisted of the Callfriend database, the full-conversations of NIST LRE 2005 and development data of NIST LRE 2007. For Russian data we used also RuSTeN (LDC 2006S34 ISBN 1-58563-388-7). VOA models are obtained from speech segments (minimum length 30 s.) extracted from VOA2 and VOA3 long files (except manually labeled files, used for testing) using telephone labels distributed by BUT.

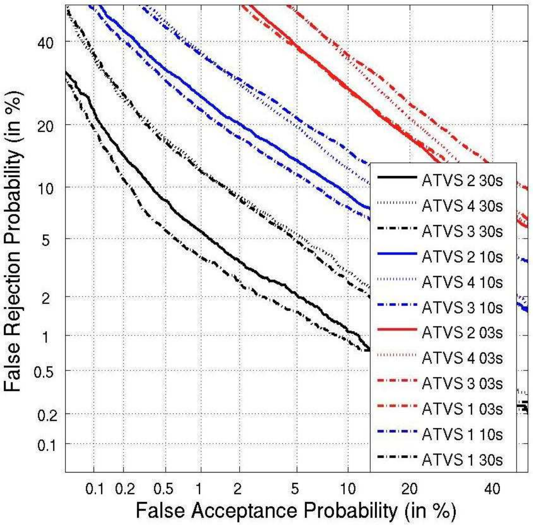

Figure 6: *Pooled DETs (EERs in %) of submitted systems on ATVS-DevTest09.*

The test material (ATVSDevTest) is obtained from LRE07Test (for target languages in both LRE07 and LRE09), and from manually labelled data from VOA2 and VOA3. A maximum of 600 test files per language is tried (600x23=13800), but finally just 5160 files were possible to extract (languages not balanced, but a much better balanced was obtained relative to DevLRE07).

For development results, see the following figures and bar charts.

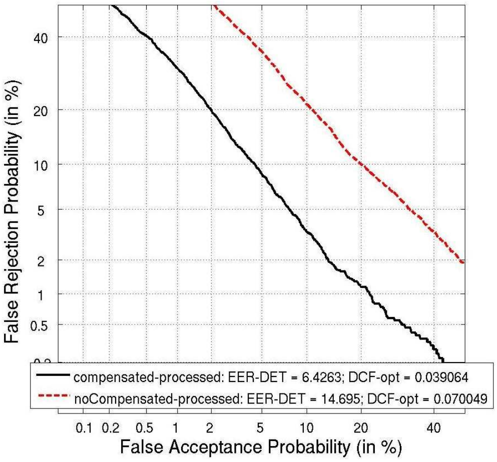

Figure 7: *Pooled DETs (EERs in %) with acoustic dot-scoring system with/without FA channel compensation on ATVS-DevTest09 (raw scores) 30s test segments.*

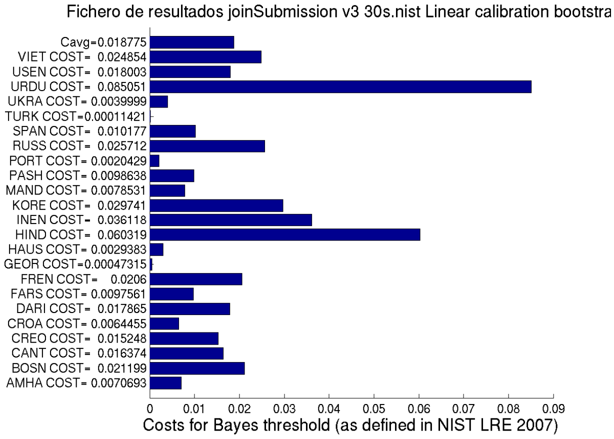

Figure 8: *Costs for Bayes thresholds of ATVS1 on ATVS-DevTest09 set for 30s test segments.*

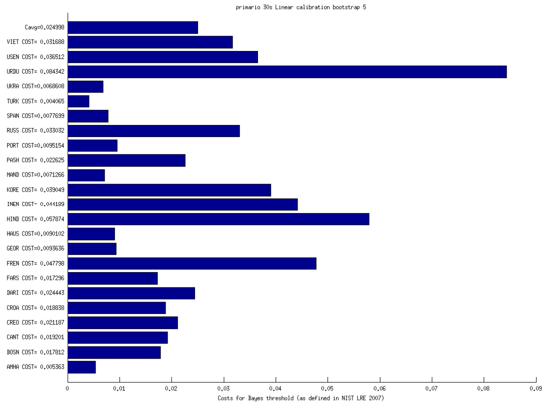

Figure 9: *Costs for Bayes thresholds of ATVS2 on ATVS-DevTest09 set for 30s test segments.*

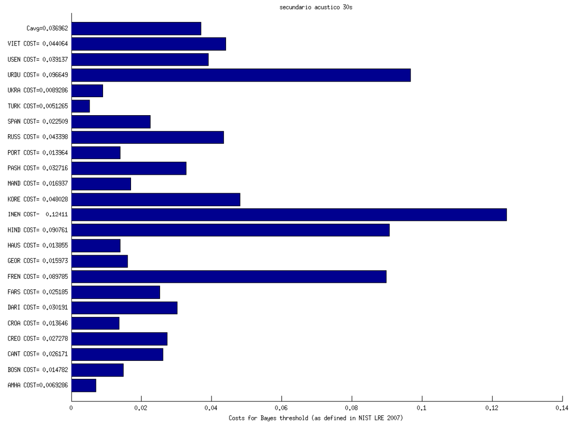

Figure 10: *Costs for Bayes thresholds of ATVS3 on ATVS-DevTest09 set for 30s test segments.*

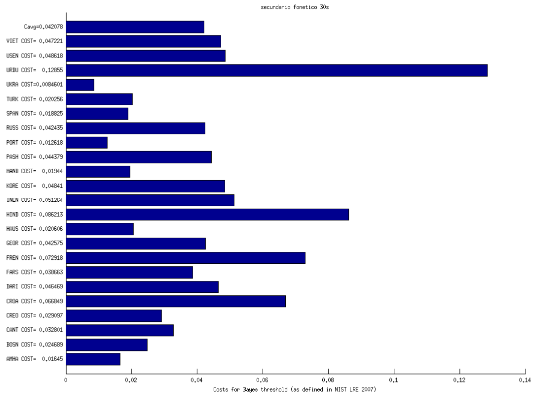

Figure 11: *Costs for Bayes thresholds of ATVS4 on ATVS-DevTest09 set for 30s test segments.*

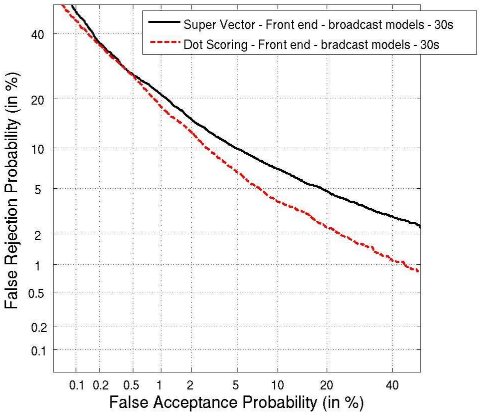

Figure 12: *Comparison of acoustic front-ends on ATVS-DevTest09 set for 30s test segments.*

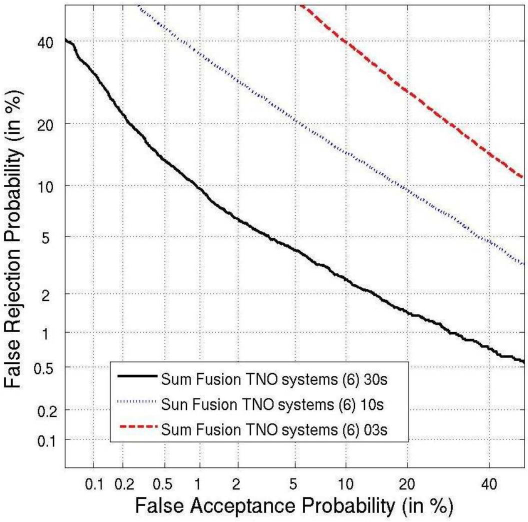

Figure 13: *Pooled DETs of TNO systems on ATVS-DevTest09 30s test segments compensation for different durations of the test segment (30s-10s-3s).*

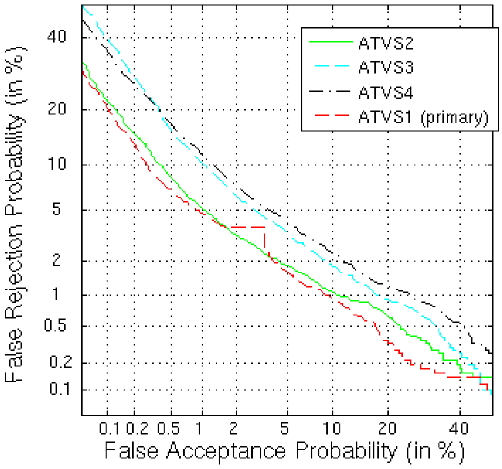

Figure 14: *Pooled DETs of calibrated submitted systems on ATVS-DevTest09 set for 30s. test segments.*

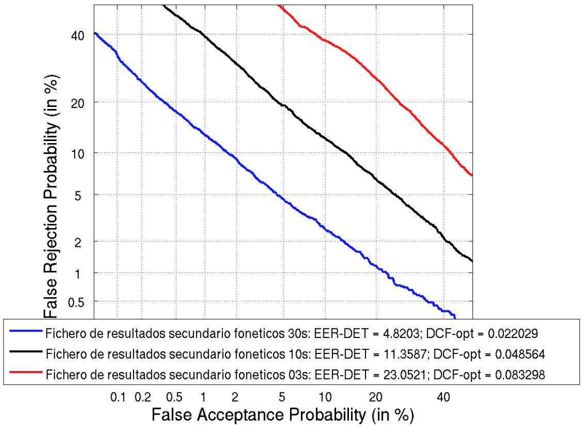

Figure 15: *Pooled DETs (EERs in %) with anchor model fusion of 10 (7 ATVS + 3 BUT) phonotactic systems (Phone-SVM) on ATVS-DevTest09 set for different durations of the test segment (30s-10s-3s).*

4. References

[Moreno 93] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Mariño, C. Nadeu, "ALBAYZÍN Speech Database: Design of the Phonetic Corpus," in proceedings of the 3rd European Conference on Speech Communication and Technology (EUROSPEECH). Berlin, Germany, 21-23 September 1993. Vol. 1. pp. 175-178.

[Campbell 06] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," Computer Speech and Language, vol. 20, no. 2-3, pp. 210–229, 2006.

[Collet 05] Mikael Collet, Yassine Mami, Delphine Charlet, Frederic Bimbot, "Probabilistic Anchor Models Approach for Speaker Verification", in INTERSPEECH 2005.

[Noorl 06] Elad Noor1, Hagai Aronowitz "Efficient Language Identification using Anchor Models and Support Vector Machines", in Odyssey 2006 ISBN: 1-4244-0472-X pp 1-6.

[SRE08] "The 2008 NIST speaker recognition evaluation http://www.nist.gov/speech/tests/spk/2008/."

[Strasheim 08] Albert Strasheim and Niko Brümmer, "SUNSDV system description : NIST SRE 2008"

[Kenny 05] Kenny, P. and Boulianne, G. and Dumouchel, P., "Eigenvoice Modeling With Sparse Training Data", IEEE Trans.~on Speech and Audio Processing, vol. 13, no. , pp 345-354.

[Vogt 08] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," Computer Speech & Language, vol. 22, no. 1, pp. 17–38, 2008.

[Lopez 08] I. Lopez-Moreno, D. Ramos, J. Gonzalez-Rodriguez and D. T. Toledano, "Anchor-model fusion for language recognition", in *Proceedings of Interspeech 2008*, Brisbane, Australia, September 2008.

Annex 1: Processing Speed

Total eval time: 620721 s.

	Ph1-Ph7 (ATVS)		Ph8-Ph10 (BUT)	
	CPU-time	Speed		Speed
parameterization	24600	25,23		
Tokenization	407400	1,52	1158560	0,54
Lattices formatting			47440	13,08
ngrams	52500	11,82	136320	4,55
Test	16000	38,80	6855	90,55
Tnorm (front-end)	70	8867,44	30	20690,70
Fusion + tnorm (back-end)	15	41381,40	15	41381,40
TOTAL	500585	1,24	1349220	0,46

ATVS+BUT PhoneSVMs CPU-TIME (sec.)	1849805
ATVS+BUT PhoneSVMs SPEED	0,335

	DS-CS (M=1024)		SV-CS (M=512)	
	CPU-time	Speed	CPU-time	Speed
Parameterization	15559,4064	39,89	15559,4064	39,89
Top-5 scoring	73245,078	8,47	26153,0448	23,73
Stats	21270,0396	29,18	19863,072	31,25
Channel compensation	334,344	1856,53	334,344	1856,53
Dot scoring	136,77	4538,43		
SVM scoring			2110	294,18
Tnorm (front-end)	10	62072,10	10	62072,10
Fusion + tnorm (back-end)	14	44337,21	14	44337,21
TOTAL	110569,638	5,61	64043,8672	9,69

TNO CPU-TIME (total, all systems)
399360

	ATVS1		ATVS 2		ATVS 3		ATVS 4	
	CPU-time	Speed	CPU-time	Speed	CPU-time	Speed	CPU-time	Speed
Front end	2407884,755	0,26	2008524,755	0,31	158719,755	3,91	1849805	0,34
formatting	4500	137,94	4800	129,32	920	674,70	3900	159,16
test	85	7302,60	90	6896,90	80	7759,01	85	7302,60
tnorm	10	62072,10	10	62072,10	10	62072,10	10	62072,10
total	2412479,755	0,26	2013424,755	0,31	159729,755	3,89	1853800	0,33