Human Assisted Speaker Recognition Using Forced Alignments on HMM

Pedro Univaso [1] Miguel Martínez Soler [1,2] Jorge A. Gurlekian [1]

[1]LIS, INIGEM,UBA-CONICET, Buenos Aires, Argentina
[2]Facultad de Ingeniería, Universidad Austral, Pilar, Argentina

Abstract

This work proposes a method for human assisted speaker recognition using an ASR system based on HMMs. Manual transcriptions are first marked at the word level and then coded by an automatic phonetic transcriptor. An initial forced alignment is made using a speaker-independent model. After this a second forced alignment is performed using each speaker-adapted model. Phonetic log-likelihood ratios are obtained and combined to get an overall score for each test. Evaluation is performed on Argentine-Spanish voice samples from the Speech_Dat database recorded on a fixed phone environment. Different recording sessions and channels for the test segments are employed. Results show a 25.1% equal error rate reduction relative to a GMM baseline system. We have used this approach for the 2012 HASR evaluation, producing three false alarms and seven misses on the twenty most difficult pairs. The proposed method could be appropriate to use in forensic tasks, where real time processing is not required.

1. Introduction

Speaker recognition main application fields are found in forensics and in security systems. Speaker identification as seen in forensics is initiated from voice recordings produced at a criminal situation. These recordings are named dubitable or evidence, and they are later matched with recordings called indubitable or suspicious that belong to a known person.

The aim of this job is to show how automatic speech recognition technology can be used to help the forensic speaker recognition task. In order to do so, we propose to use an ASR system to perform forced alignments, given human transcriptions marked at the word level. We use the phoneme scores from the alignments as partial scores that are averaged to get a final score. Since we are dealing with speaker recognition, we get scores in a likelihood-ratio framework, taking a score from a UBM model (i.e.: a speaker-independent speech recognition model) and a target speaker model (i.e.: a speaker-dependent speech recognition model).

This way of dealing with the problem, configures a human assisted approach. Since NIST introduced the HASR test as a pilot evaluation there has been some research in the area, most of it related to the way humans can complement automatic speaker recognition systems (as in [1]) or simply to evaluate the ability of humans in a speaker recognition task (as in [2] and [3]).

However, we have no knowledge about a system that tries to use human speech recognition ability to provide useful information to a speaker recognition system. That is the reason for trying to evaluate that possibility in the present work.

The rest of this paper is as follows: Section 2 presents the proposed method based on phonological information. Section 3 shows the strategy of evaluation. Section 4 presents the results obtained, which are discussed in section 5, to finally present our conclusions and future work in sections 6 and 7, respectively.

2. Methodology

The proposed methodology consists in the extraction of acoustic information from phonemes by forced alignment performed with an ASR system. Prior to the automatic alignment, manual transcriptions are first marked at the word level and then coded by an automatic phonetic transcriptor. Grapheme to phoneme conversion is a key component for applications involving speech recognition and synthesis. Unlike English, Spanish has a fairly consistent mapping from graphemes to phonemes. Due to this characteristic, it is possible to build automatic converters based on rules. The automatic phonetic transcriptors used in this work are two. The first one is based on a cascade of rules, manually designed, and tested using manually annotated corpus for Argentine-Spanish [13]. The second one relies on a dictionary approach and is applied to the English tests in the HASR evaluation.

The motivation for using a speaker recognition system based on forced-alignment is that a human can make word and phonetic transcripts in a forensic environment, both for the known and unknown speaker samples that are going to be tested.

2.1 GMM Models

A GMM model can be viewed as an HMM with a unique single state. We implemented our GMM model as an HMM with that single state, but adding to it a silence model. This silence model is not used in the decision stage (i.e.: no scores are computed on them), but it is useful as a voice activity detector.

Our reference GMM model is based on the Universal Background Model approach presented by Reynolds et. al. in [5]. Bayesian adaptation via maximum a posteriori estimation (MAP) is used to adapt speaker models from the UBM. A simplified schematic of the speaker recognition system based on the GMM-UBM methodology can be seen in Fig. 1.

This model makes no use of the manual transcriptions and its only purpose is to serve as a baseline system.

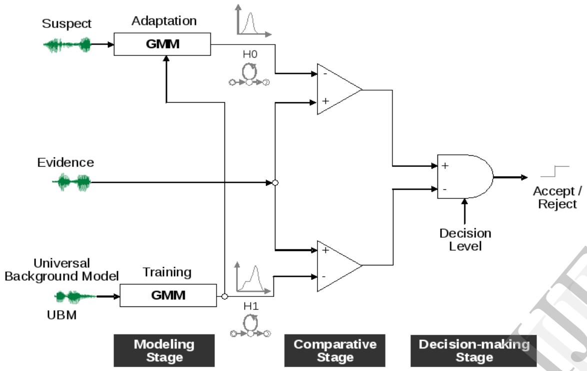

Figure 1. GMM-UBM methodology

2.2 Hidden Markov Models

Fig. 2 shows the procedure used in this study to recognize speakers using an HMM-based ASR system. The generation of a universal reference (UBM) speaker acoustic model was performed by a training process in stages according to the methodology proposed by Young [6] for ASR systems. HMMs are usually built to model either single isolated phones or context-dependent triphones. We chose the latter approach in this work.

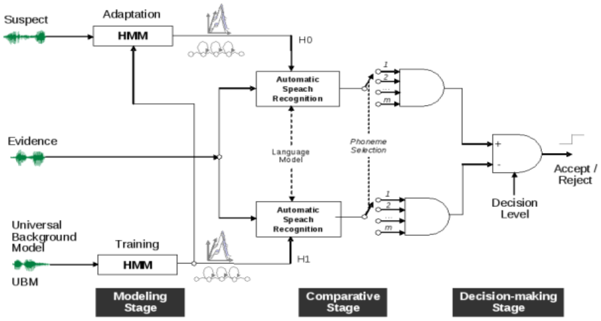

Figure 2. HMM-UBM methodology.

Target model adaptation was conducted similarly to the case of GMM, re-estimating each HMM triphone model with information of the target speaker. We employ forced alignment to get scores from model-segment pairs, one comparing the UBM model with the test segment, and the other comparing the target adapted model with the test segment. In both of them, a Viterbi algorithm uses a word network to align human made transcriptions.

In the final stage of comparative analysis, we calculate average log-likelihoods for each aligned phoneme to finally obtain overall average values for each model-segment pair comparison. These two values show how well both the UBM model and target adapted model explain the data observed in the test segment. A final score is computed as a log-likelihood ratio between both model scores.

However, we have three scoring strategy variations for computing the average log-likelihoods that was mentioned in the previous paragraph. The first one consists in normalizing phone scores by the number of scores that appeared in the test utterance. This corresponds to the HMM method in the tables below. The second one consists in computing an average score for each phone that appeared in the sentence and after that, normalizing by the number of phones that appeared in the sentence. We call this approach HMMphn in further tables. Finally, the last one (denoted DHMM), consists in computing taking into account each phoneme in accordance with its discrimination factor. The discrimination factor (Table 1) was calculated based on the log-likelihood-ratio cost function (C_{llr}) [9] using the HMM method, with a phoneme at a time, over the training data.

Table 1. Phoneme discrimination ranking based on C_{llr} and DHMM Discrimination Factor (Dfi).

Rnk	Phoneme	C_{llr}	Dfi
1	n	0,58	3,5
2	t	0,58	3,4
3	o	0,60	3,3
4	s	0,81	2,5
5	m	0,93	2,2
6	N	0,93	2,2
7	k	0,94	2,1
8	J	1,00	1,0
9	R	1,00	1,0
10	Z	1,02	1,0
11	f	1,04	1,0
12	g	1,04	1,0
13	u	1,04	1,0
14	H	1,08	1,0
15	x	1,08	1,0
16	r	1,14	1,0
17	h	1,15	1,0
18	a	1,21	1,0
19	w	1,23	1,0
20	B	1,25	1,0
21	i	1,31	1,0
22	p	1,32	1,0
23	e	1,39	1,0
24	d	1,42	1,0
25	l	1,44	1,0
26	j	1,61	1,0
27	G	1,64	1,0
28	D	1,76	1,0
29	b	1,82	1,0
30	C	1,99	1,0

The final score, using the phoneme discrimination factor, was computed with the following equation:

$$LL = \frac{1}{\sum_i Dfi} \cdot \sum_i \left(Dfi \cdot LLi \right) \qquad (1)$$

Where *LLi* is the log-likelihood considering only the *i* phone, and *Dfi* is the discrimination factor for the *i* phone. The discrimination factor takes the following values:

$Dfi = 2/C_{llr}$ if $C_{llr} < 1$
$Dfi = 1$ if $C_{llr} \geq 1$

All these values were computed using the training corpus data.

3. Evaluation

Two different experiments will be presented in this section. The first one is a speaker recognition test that was performed using the proposed methodology on an Argentine-Spanish database augmented with recent recordings of a subset of speaker in mismatched condition. The second experiment was held during the NIST Human Assisted Speaker Recognition evaluation that took place as a part of NIST SRE 2012 evaluation.

3. Database

The database used in this work is part of the SALA I Project (SpeechDat Across Latin America) [7]. The style of speaking corresponds to read paragraphs taken from newspapers and books of Argentina or developed by linguists. Recordings were made through the fixed telephone network through a computer equipped with an AVM-ISDN-A1 board and a basic access interface ISDN (BRI).

The SALA I corpus [8] was divided in five dialectal regions. We used utterances from the South region in order to build the UBM model and target speaker models. This training corpus comprises 1,301 utterances, with a total of 9,948 words, corresponding to a vocabulary of 2,722 different words, issued by 136 speakers (47 males and 89 females) for an overall 99 minutes of recording. From the total 136 speakers, 130 were used to build the UBM model and the remaining 6 speakers to build target models.

For the testing corpus, we localized six subjects (3 male and 3 female) that were present in the SALA I recordings that took place eleven years ago. They were recorded with the same protocol used in SALA I, but using a direct laptop microphone rather than a phone channel. This testing corpus contains 4 utterances for each speaker, leading to 24 utterances for the whole set. These 24 segments are compared to the six target models, resulting in 144 test pairs (24 target and 120 non-target).

3.1 Implementation

Audio segments were coded at 8 kHz, 16 bit. Signal mean subtraction was used to eliminate any offset from the analog recording stage. We employed a 25 ms Hamming window at a 10 ms rate, a pre-emphasis filter (coeff=0.97), and energy normalization to get 13 MFCC coefficients with delta and acceleration. These coefficients were used as input features to train GMM a HMM models on HTK Toolkit ver. 3.4 [6].

4. Results

Table 2 shows comparative results of equal error rate, Cmin$_{llr}$ and C$_{llr}$ for tests performed using the original waves without filtering. Table 3, on the other hand, shows the same error metrics using test segments that were passed through a G.712 filter [4]. This filter simulates the transmission characteristics of pulse-code modulation channels and was used for channel compensation since the training data was produced over a phone channel while the test on a direct laptop microphone. The performance improvement after filtering is apparent.

The use of zero normalization or Z-Norm, another method for handling channel mismatch conditions that have been mainly applied to verification, resulted in a worse performance than filtering, and so these results are not shown in this paper.

In order to evaluate the effectiveness of the proposed approach an experiment with a baseline HMM system without manual transcription (i.e. a typical ASR system) was performed. A C$_{llr}$ of 1.34 (vs. 1.24) for the original waves and a C$_{llr}$ of 0.84 (vs. 0.78) for the filtered ones show that forced-alignment produces an average improvement of 7% over an ASR baseline with a word recognition accuracy of 71.51%.

Since no calibration was performed, we will analyse differences among systems assessing discrimination error by means of the Cmin$_{llr}$ metric. However, C$_{llr}$ was included also to show how the mismatched condition affects the calibration error.

Table 2. Equal error rate (EER), Cmin$_{llr}$, and C$_{llr}$ for the experiments on SALA I, testing on mismatched condition without filtering.

Method	EER (%)	Cmin$_{llr}$	C$_{llr}$
GMM	25.0	0.62	1.25
HMM	20.8	0.51	1.24
HMMphn	20.8	0.51	1.19
DHMM	20.8	0.49	1.09

Table 3. Equal error rate (EER), $Cmin_{llr}$, and C_{llr} for the experiments on SALA I, testing on mismatched condition after passing a G.712 filter.

Method	EER (%)	$Cmin_{llr}$	C_{llr}
GMM	16.7	0.57	0.87
HMM	16.7	0.38	0.78
HMMphn	12.9	0.38	0.77
DHMM	12.5	0.38	0.70

4.1. HASR 2012

The Human Assisted Speaker Recognition 2012 evaluation was a pilot evaluation that took place during the NIST 2012 Speaker Recognition Evaluation as an attempt to address the question of how humans can effectively interact with automatic speaker recognition technology. Participants were allowed to combine automatic processing with human involvement. To accommodate different interests and levels of effort, two test sets were offered by NIST, one with 20 trials (HASR1), and one with 200 trials (HASR2).

Each test consisted in comparing a speaker in a telephone conversation excerpt with another one in an interview segment, recorded in high quality. Trials were delivered one at a time. A participant had to submit a score for the first trial in order to get the second one, then submit a score for that one in order to get the third, and so on. We chose the short HASR1 protocol, which contained the 20 most difficult pairs, since the time complexity of Viterbi algorithm made impossible to process 200 tests in time for the submission deadline.

The LIS submission to HASR 2012 consisted in scores produced by a system that implements the method described in this paper, adding manual wave edition in order to filter out bad quality segments (i.e. sections with noise, non-speech activity, other speaker voices, etc.). Since we did not have an American English labelled data base, we used an open source speaker independent ASR model from Carnegie Mellon University [11] as our UBM (trained on Communicator Corpus [12]). We used this model to adapt it to the target speaker utterances by multi-class MLLR. The phonetic classes used for multi-class MLLR adaptation were the same as in [10]. A set of impostor models was also adapted to perform T-Score normalization. The wave segments used as impostors were taken from the 2008 NIST 10sec training data and short3 test data, taking excerpts in English only.

Since we needed phoneme transcriptions to perform normalization, we produced 56 non-native human transcripts for 10sec segments taken randomly. The normalization set labels were completed using ASR transcripts provided by a similar number of short3 segments.

Scores were computed as it was described previously in the methodology section for the HMMphn system, but selecting scores from vowel, diphthong, nasal and liquid phones only, since they showed to be more robust when tested in a development data set. After that, T-Norm was applied to produce the final scores. The whole process was run on a slightly modified version of the CMU Sphinx3 engine that does not perform score scaling.

Results for this system on HASR 2012 data are shown on Table 4. In our original submission, we did not apply any filtering to the interview segments of HASR1 trials. We got three false alarms and seven misses on those 20 pairs using a threshold determined using a development set. However, after performing the experiments described above, we applied the filter to the 20 tests in order to measure whether there is an improvement in performance or not.

Table 4. Equal error rate, $Cmin_{llr}$, and C_{llr} for the experiments on HASR1 data, testing on mismatched condition before and after passing a G.712 filter.

Condition	EER (%)	$Cmin_{llr}$	C_{llr}
Original	65.0	0.82	1.10
G.172 filter	45.0	0.72	1.10

5. Discussion

Results of HMM forced alignment show the best performance among all experiments. The use of forced alignments guided by a human proved in this work a clear improvement over a GMM model. This results support the idea that forensic speaker recognition could benefit from this basic idea of using the known text to produce forced alignments and then to continue with an automatic recognition paradigm using a universal background model of a language region. Moreover, forensic scientists could select portions of the test segment to be used, given some standard procedure (i.e.: standard metrics for assessing quality, energy level, presence of other speaker, etc.).

However, it must be noticed that in order to successfully implement a speaker recognition system of this type, we need to add some session compensation mechanism which proves to be successful in eliminating inter-session variability. In this sense, the drop in performance that we got due to mismatched condition, even after filtering, tells us that there is some variability that was not modeled. In this respect, the mismatch due to aging must have contributed significantly.

Regarding the HASR1 tests, the $Cmin_{llr}$ metric shows that the system had some discrimination power, despite the poor EER. The situation improved after filtering the interview segments, however. Considering that the transcripts were made by non-native speakers, we believe that the poor performance could be explained, at least in part, by mistakes made during the human transcribing process.

In summary, the human assisted speaker recognition approach using forced alignment presents results that are highly promising and should be tested on bigger corpora.

6. Conclusions

Determining the most efficient way to use speaker intrinsic information to improve speaker identification is still a debt we have today. The methodology presented here is an alternative approach that attempts to contribute to solve this issue. The way phonemes are produced by the speaker and their variations are highly speaker dependent.

Modelling speaker voice using HMM contributes to enhance phonological information present in the speaker model. The use of forced alignment in HMMs is possible for forensic purposes, where the court requests the use of speaker identification techniques using transcriptions made in advance. We think that Human Assisted approaches are the future of forensic speaker recognition, since in a framework like that, the expert's work can be divided into a manual evidence acquisition stage, and an automatic decision making, providing both transparency and flexibility.

7. Future work

We are building up a new multi-channel and multi-session database of Argentine Spanish that will be used to test the method in channel and session mismatch conditions. In that sense, we plan to add a Total Variability model based on HMM model means and a PLDA model for inter-session variability compensation.

Acknowledgments

To project PID 35891, "Development of techniques for speaker recognition", awarded by the Ministry of Science, Technology and Productive Innovation of Argentina.

8. References

[1] R. Schwartz, J. Campbell, W. Shen, D. E. Sturim, W. M. Campbell, F. S. Richardson, R. B. Dunn et al. *USSS-MITLL 2010 human assisted speaker recognition*. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference, pp. 5904–5907 (2011).

[2] D. Ramos, J. Franco-Pedroso, J. Gonzalez-Rodriguez, "Calibration and weight of the evidence by human listeners. The ATVS-UAM submission to NIST HUMAN-aided speaker recognition 2010". *Acoustics, Speech and Signal Processing (ICASSP),* IEEE International Conference, pp. 5908–5911 (2011).

[3] W. Shen, "Assessing the Speaker Recognition Performance of Naive Listeners Using Mechanical Turk", *Contract,* pp. 5916–5919 (2010).

[4] H. G. Hirsch, "FaNT - Filtering and Noise Adding Tool" (2005)

[5] D. Reynolds, T. Quatieri, R. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, 10(1-3), 19-41 (2000)

[6] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtech, V. Wooland, P.: The HTK Book. Cambridge University Press (2006).

[7] A. Moreno, "SALA: SpeechDat Across Latin America", *Proceedings of the I Workshop on Very Large Databases*, Athens, Greece (2000).

[8] J. A. Gurlekian, L. Colantoni, H. Torres, A. Rincón, A. Moreno, J. Mariño, "Database for an automatic Speech Recognition System for Argentine Spanish", *Proc. of the IRCS Workshop on Linguistic Databases*, pp. 92-98, Pennsylvania (2001).

[9] N. Brümmer and J. du Preez, "Application-Independent Evaluation of Speaker Detection", *Comput. Speech Lang.*, 20(2-3), 230-275 (2006).

[10] A. Stolcke, SS. Kajarekar, L. Ferrer and E. Shriberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms", *IEEE Trans. Audio, Speech, and Lang. Process*. vol. 15, pp. 1987–1998, (2007).

[11] Available at: http://www.speech.cs.cmu.edu/sphinx/models/communicator_mar2008/communicator_4000_20080321.tar.gz.

[12] C. Bennett, and A. Rudnicky, "The Carnegie Mellon Communicator Corpus", *Proceedings of ICSLP*, Denver, Colorado. Retrieved from http://repository.cmu.edu/compsci/1395/ (2002).

[13] D. Evin, "Grapheme to Phoneme Conversion for Argentine-Spanish". Technical Report of LIS - Laboratorio de Investigaciones Sensoriales, Argentina, 2009.