X-vectors based Urdu Speaker Identification for short utterances

Muhammad Umar Farooq, Farah Adeeba, Sarmad Hussain
Center for Language Engineering,
Al-Khawarizimi Institute of Computer Sciences,
University of Engineering and Technology, Lahore, Pakistan.
{umar.farooq, farah.adeeba, sarmad.hussain}@kics.edu.pk

***Abstract*—In context of commercial applications, robustness of a Speaker Identification (SI) system is adversely effected by short utterances. Performance of SI systems fairly depends upon extracted feature sets. This paper investigates the effect of various feature extraction techniques on performance of i-vectors and x-vectors based Urdu speakers' identification models. The scope of this paper is restricted to text independent speaker identification for short utterances (up to 4 seconds). SI systems demand for a large data covering sufficient inter-speaker and intra-speaker variability. Available Urdu speech corpus is used to measure performance of various feature sets on SI systems. A minimum percentage Equal Error Rate (%EER) of 0.113 is achieved using x-vectors with Linear Frequency Cepstral Coefficients (LFCCs) feature set.**

***Index Terms*—speaker identification, deep neural networks, speaker embeddings, i-vector**

I. Introduction

Speaker Identification (SI) is the task of identifying a person, based on given speech signal and enrolled speaker record [1]. If the lexical content of the utterance is fixed to some phrase, the task is considered as text-dependent, otherwise it is text-independent. Speaker identification is widely applied for speaker surveillance, forensics, multi-speaker tracking and speaker authentication [2]. Introduction of subspace modeling techniques such as Joint Factor Analysis (JFA) [3] and i-vector [4] has made tremendous progress in the field of text-independent speaker identification [5]. With increasing trend of commercialization, many applications require very good accuracy even with short duration utterances. However, the performance of JFA and i-vector degrades with short utterances of about 5-10 seconds [6].

I-vector is the widely used speaker modeling technique for speaker identification systems. It consists of a pipeline of generative models, trained on independent subtasks: a Universal Background Model (UBM), a total variability matrix *(T)* to extract i-vetcors and PLDA [7] backend to compute similarity score between i-vectors [4] [8]–[12]. UBM, a Gaussian Mixture Model (GMM), is trained using very large data to collect sufficient statistics for training of i-vector extractor. UBM is mostly trained using iterative Expectation-Maximization (EM) algorithm.

978-1-7281-2449-0/19/$31.00 ©2019 IEEE

Recently, Deep Neural Network (DNN) based end-to-end speaker identification is introduced that learns speakers' embeddings [13]. In order to make all segments of same length, Snyder et al. [14] introduced a temporal pooling layer in network for length normalization. The work in [15] splits this end-to-end training into two parts such as a DNN to produce embeddings and a separately trained classifier to compare these embeddings. This facilitates the use of all the accumulated backend technology developed over the years for i-vectors, such as length-normalization, PLDA scoring, and domain adaptation techniques.

In addition to the speaker modeling techniques, feature extraction plays vital role in speaker identification process since accuracy of the system fairly depends upon selection of the acoustic characteristics that maximize discrimination between the speakers. Mel Frequency Cepstral Coefficents (MFCC) [16], Linear Frequency Cepstral Coefficents (LFCC), Gammatone Frequency Cepstral Coefficents (GFCC) [17], filter bank (fBank) [18] coefficients and Perceptual Linear Predictive (PLP) [19] are the widely used feature extraction techniques for SI systems. In this paper, comparative analysis of these feature sets on speaker modeling is performed. UBM based i-vectors and state-of-the-art DNN based x-vectors are used for speaker modeling.

Rest of the paper is organized as follows: Section II briefly describes the components of a SI system, Section III explains the data set used for training, enrollment and evaluation of SI systems. Experimental setup is described in Section IV and Section V depicts results of i-vectors and x-vectors with various feature sets.

II. Speaker Identification System

A speaker identification system consists of three basic components that include feature extraction, speaker modeling and scoring trials which is shown in Figure. 1.

A. Feature extraction

MFCCs are the widely used features in most of the speech processing applications such as speech recognition [20], language identification [21] and speaker identification [22]. Key steps involved in MFCCs extraction are shown in Figure 2.

In addition to MFCC features, Mel filter-banks features are also used in current work. Mel filter-banks are the array of pass

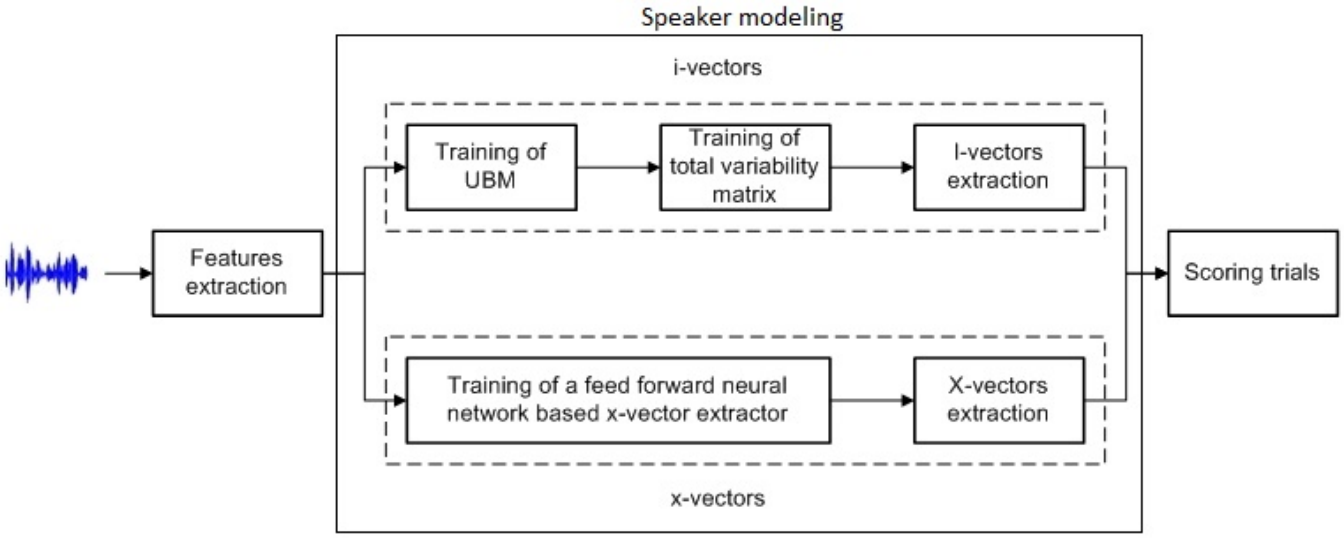

Fig. 1. Components of i-vector and x-vector based SI system

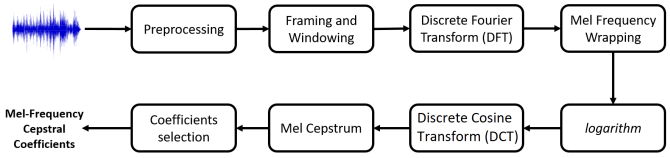

Fig. 2. Flow of MFCC extraction

band filters that separates an input acoustic signal into multiple components [18]. The key difference between MFCC features and Mel filter-bank features is the truncation.

Mel filter bank focuses on lower frequencies region as compared to the higher ones. However, based on speech production theory, speaker characteristics are more prominent in higher frequency region of speech [23]. Linear filter bank emphasizes all frequencies equally. This insight motivates to investigate performance of LFCCs for speaker recognition.

GFCC features are another type of acoustic features used for speaker identification system. Extraction of GFCC is just similar to calculating MFCCs except the bandwidths of filter banks applied for frequency wrapping. GFCCs are extracted after applying Gammatone Filter (GF) bank for frequency wraping. GF is a linear filter described by an impulse response which is the product of a sinusoidal and a gamma distribution.

PLP is another widely used acoustic feature extraction technique [19]. Power spectrum from speech signal is computed and bark filter bank is applied. These filter banks are weighted by equal-loudness pre-emphasis weights to simulate hearings sensitivity. A linear prediction is applied to wrapped spectrum to predict coefficients of a signal that has this wrapped spectrum as a power spectrum. Finally, cepstral coefficients are obtained from these linear predicted coefficients.

B. Speaker modeling

1) I-vector: Joint Factor Analysis (JFA) [3] based SI systems are outperformed by i-vector. JFA separates a Gaussian supervector as a sum of speaker and channel supervectors. However, studies show that channel factors also have speaker information [24]. This problem motivated the introduction of i-vector that maps both speaker and channel variability into same low-dimensional space. A supervector M of a certain utterance can be split as;

$$M = m + Tw \tag{1}$$

where m is the vector of speaker independent components, T is the total variability matrix and w is known as i-vector. Once sufficient parameters given in Equation 1 are learned, i-vector for an input utterance can be extracted. A UBM is used to collect sufficient statistics to train i-vector extractor. Different number of mixtures with various i-vector dimensions are experimented to optimize i-vector performance.

2) X-vectors: X-vector is state-of-the-art speaker modeling technique which is based on DNN speaker embeddings [15]. These embeddings are extracted from a feed forward deep neural network. Network is consisted of some initial frame-level layers, a statistics pooling layer and a few segment-level layers. Statistic pooling layer aggregates frame-level representations and calculates its mean and standard deviation. These segment-level statistics are concatenated together and passed to segment-level hidden layers. Embeddings can be extracted from these layers. In this paper, x-vector system described in [25] is investigated using various feature sets.

C. Scoring

PLDA [7] is used to score similarity between two i-vectors or x-vectors. Since vectors are modeled by a factor analyzer in PLDA, it outperforms conventional cosine distance scoring

[26]. Dimmenisionality reduction is done using Linear Discriminant Analysis (LDA) [15]. PLDA model is trained using large in-domain data.

III. Data set

For development and evaluation of speaker identification system, a speech corpus collected for Urdu LVCSR [20] is used. Most of the speakers are recorded in multiple sessions hence covering the session variability. Data is recorded from male and female Punjabi and Urdu speakers from age group ranging between 18-50 years. All audios are recorded on a sampling rate of 16KHz using USB microphone, USB headsets, hands-free and laptop microphone. Data is recorded in indoor environment. Each speaker is asked to record sentences from Urdu text corpus discussed in [20]. Speech corpus is split into train, enrollment and test sets. Train set consists of 1575 speakers and is used to train UBM and PLDA. For enrollment and test sets, speech corpus from 300 speakers is selected. Enrollment and test data sets are restricted to speech duration of 4 minutes and 1 minute per speaker respectively. Additionally, each utterance in both sets is less than 4 seconds. Details about speakers and utterances duration are shown in Table I.

TABLE I
Speakers' details

	Train	Enrollment	Test
No. of speakers	1575	300	300
Shortest utterance (s)	1.02	1.30	1.39
Longest utterance (s)	34.75	3.99	3.99
Mean duration (s)	5.07	3.26	3.35
Total utterances	198594	13156	5523
Total duration per speaker (s)	640[1]	240	60

IV. Experimental setup

Different feature extraction techniques are used to study effect of various feature sets on speaker identification system. Performance of i-vector and x-vectors are compared using MFCCs, LFCCs, GFCCs, fBanks and PLP features. Features are extracted using a 25ms hamming window with a shift of 10ms. 13 coefficients are calculated for all feature extraction technique except filter banks where 24 filter banks are applied. A Linear Predictive Coding (LPC) filter of order 12 is used for PLP extraction. Summary of feature sets is given in Table II . These features are used to train UBM, i-vector and x-vector extractors.

To train i-vector extractor, a full covariance GMM model (UBM) is trained using train set. Performance of speaker identification system varies with number of mixtures in UBM and dimensionality of i-vector. Various combinations of Gaussian mixtures and i-vector dimensions are investigated to optimize the performance. PLDA scores of test trials are evaluated for each combination and the best %EER is reported for comparison.

[1]Average duration per speaker in train set

TABLE II
Number of coefficients calculated for different feature sets

Feature type	No. of coefficients
MFCC	13
LFCC	13
GFCC	13
fBank	24
PLP	13

For proper training of x-vectors, sufficient duration of each speaker and utterance length is required. So, all utterances with less than 5 seconds and speakers with less than 8 utterances are discarded from train set for x-vector extractor training. For neural network training, different cell dimensions are investigated with a 6-layers network proposed by [25]. First 5 layers are frame level with time-delay architecture. Next layer is statistics pooling layer followed by two segment level layers. Final layer is a softmax output layer.

Different combinations of Gaussian mixtures in UBM and i-vector dimensionality are investigated to optimize the i-vector performance. Best configuration with 128 Gaussian mixtures in UBM and i-vector of dimension 600 is selected to compare performance of various feature sets. Similarly, for x-vectors extraction, various combinations of number of neurons in hidden layer, static pooling layer and number of epochs are experimented to achieve the optimal accuracy. Best configuration with 1024 neurons in each layer and 14 training epochs is used to compare performance of features sets. Kaldi speech recognition toolkit [27] is used for speaker modeling.

Equal Error Rate (EER) is the widely quoted measure to evaluate performance of a SI system. False acceptances and false rejections are calculated defining a threshold for trials' scores. EER is the value where False Rejection Rate (FRR) and False Acceptance Rate (FAR) becomes equal for given threshold. In speaker identification evaluations, trade-off between false alarms and missed speakers has always been an important diagnostic tool [28]. NIST has defined Detection Cost Function (DCF) and Detection Error Tradeoff (DET) [29] curves as a diagnostic measure. DCF is the weighted sum of probabilities of missed speakers and false alarms. DCF and EER are calculated on a certain threshold. But a single performance number is inadequate to represent capabilities of a system with multiple operating points. SI systems have many operating points and best can be observed by a performance curve. It can easily be visualized plotting a DET curve.

V. Results

Performance of i-vectors with various feature sets is shown in Table III in terms of Equal Error Rate (%EER) and minimum DCF at $P_{Target} = 0.01$. Figure 3 shows the DET curve comparing the performance of i-vectors with all investigated feature sets. Curve illustrates that LFCCs perform slightly better than other feature extraction techniques. It overlaps with MFCCs for some region.

Best configuration of x-vectors is experimented with different feature extraction techniques. Percent equal error rate

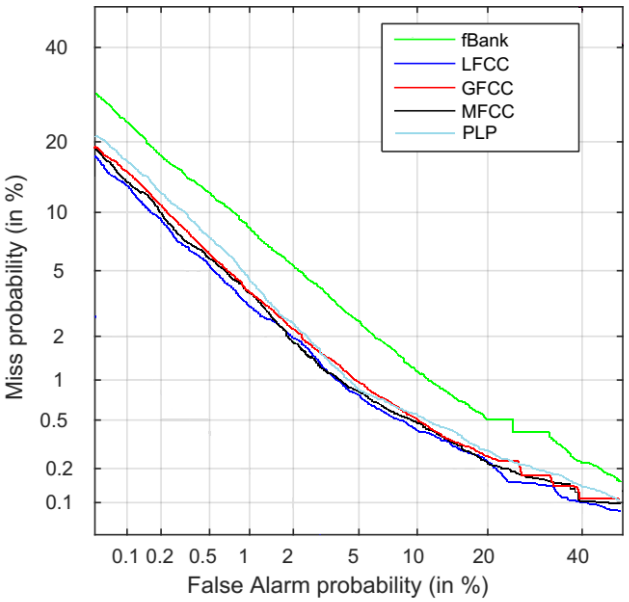

Fig. 3. DET curve comparing various feature sets for i-vectors

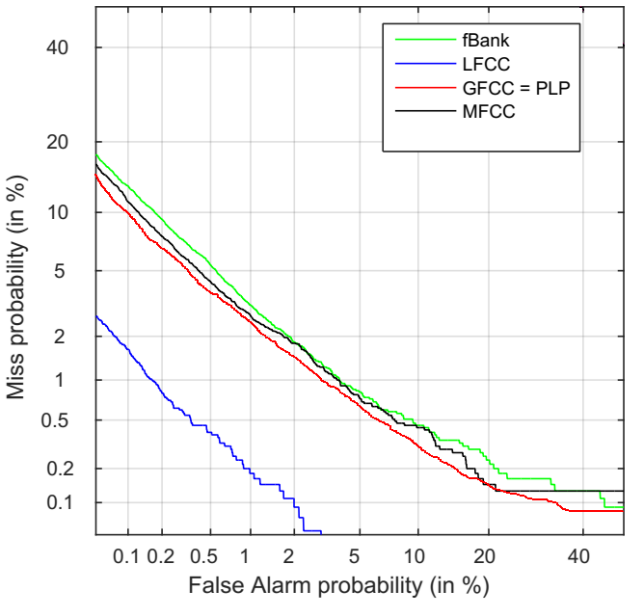

Fig. 4. DET curve comparing various feature sets for x-vectors

TABLE III
COMPARISON OF DIFFERENT FEATURE TYPES ON BEST I-VECTOR CONFIGURATION

Feature type	% EER	DCF
MFCC	0.914	0.051
LFCC	**0.701**	**0.039**
GFCC	1.167	0.081
fBank	1.449	0.102
PLP	1.209	0.067

is shown in Table IV and DET curve is shown in Figure 4. DET curve shows that LFCCs perform far better than all other feature extraction techniques for all operating points.

TABLE IV
COMPARISON OF DIFFERENT FEATURE TYPES ON BEST X-VECTOR CONFIGURATION

Feature type	% EER	DCF
MFCC	0.739	0.041
LFCC	**0.113**	**0.004**
GFCC	0.653	0.037
fBank	0.943	0.053
PLP	0.653	0.037

It is evident from experimental results that x-vectors outperform i-vectors using LFCCs which is in accordance with theoretical explanation of LFCCs covering all frequency regions equally.

VI. CONCLUSION

This paper presents the effect of different acoustic feature extraction techniques on state-of-the-art speaker embeddings extracted from neural networks to discriminate between different speakers. Urdu speech corpus collected from Pakistani speakers is used to evaluate and compare different SI systems. This work is restricted to compare text independent SI systems' performances on short duration utterances up to 4 seconds. Experiments show that state-of-the-art x-vectors outperform i-vectors. A minimum %EER of 0.113 is achieved using LFCCs which outperformed the other investigated feature sets such as GFCCs, MFCCs, fBank cepstral coefficients and PLP.

REFERENCES

[1] H. Beigi, *Fundamentals of Speaker Recognition*. Springer US, 2011.
[2] N. Singh, R. Khan, and R. Shree, "Applications of speaker recognition," *Procedia Engineering*, vol. 38, pp. 3122 – 3126, 2012, iNTERNATIONAL CONFERENCE ON MODELLING OPTIMIZATION AND COMPUTING. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187770581202276X
[3] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12 – 40, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639309001289
[4] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.
[5] T. Stafylakis, M. J. Alam, and P. Kenny, "Text-dependent speaker recognition with random digit strings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1194–1203, July 2016.
[6] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7649–7653.
[7] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 531–542.
[8] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, Oct 2007, pp. 1–8.

[9] J. A. V. López and N. Brümmer, "Towards fully bayesian speaker recognition: Integrating out the between-speaker covariance," in *INTERSPEECH*, 2011.

[10] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey*, 2010.

[11] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH*, 2011.

[12] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4257–4260.

[13] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5115–5119.

[14] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016, pp. 165–170.

[15] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *INTERSPEECH*, 2017.

[16] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *ACOUSTICS, SPEECH AND SIGNAL PROCESSING, IEEE TRANSACTIONS ON*, pp. 357–366, 1980.

[17] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 4625–4628.

[18] S. V. Chougule, M. S. Chavan, and M. S. Gaikwad, "Filter bank based cepstral features for speaker recognition," in *2014 IEEE Global Conference on Wireless Computing Networking (GCWCN)*, Dec 2014, pp. 102–106.

[19] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[20] M. U. Farooq, F. Adeeba, S. Rauf, and S. Hussain, "Improving large vocabulary urdu speech recognition system using deep neural networks," in *INTERSPEECH*, (accepted) 2019.

[21] F. Adeeba and S. Hussain, "Acoustic feature analysis and discriminative modeling for language identification of closely related south-asian languages," *Circuits, Systems, and Signal Processing*, vol. 37, no. 8, pp. 3589–3604, Aug 2018.

[22] F. Leu and G. Lin, "An mfcc-based speaker identification system," in *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*, March 2017, pp. 1055–1062.

[23] B. H. Story, "Using imaging and modeling techniques to understand the relation between vocal tract shape to acoustic characteristics," in *in Proc. Stockholm Music Acoustics Conf., 2003, SMAC-03*, pp. 435–438.

[24] N. Dehak, "Discriminative and generative approaches for long- and short-term speaker characteristics modeling: Application to speaker verification," Ph.D. dissertation, 2009, aAINR50490.

[25] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5329–5333.

[26] C. Zhao, L. Li, D. Wang, and A. Pu, "Local training for plda in speaker verification," in *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, Oct 2016, pp. 156–160.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[28] D. A. van Leeuwen and N. Brümmer, *An Introduction to Application-Independent Evaluation of Speaker Recognition Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 330–353.

[29] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The det curve in assessment of detection task performance," in *EUROSPEECH*, 1997.