An investigation of likelihood normalization for robust ASR

Emmanuel Vincent, Aggelos Gkiokas, Dominik Schnitzer, Arthur Flexer

An investigation of likelihood normalization for robust ASR

Emmanuel Vincent[1], *Aggelos Gkiokas*[2], *Dominik Schnitzer*[3], *Arthur Flexer*[3]

[1]Inria, Villers-lès-Nancy, F-54600, France
[2]Institute for Language and Speech Processing, Athens, Greece
[3]Austrian Research Institute for Artificial Intelligence
emmanuel.vincent@inria.fr, agkiokas@ilsp.gr, arthur.flexer@ofai.at

Abstract

Noise-robust automatic speech recognition (ASR) systems rely on feature and/or model compensation. Existing compensation techniques typically operate on the features or on the parameters of the acoustic models themselves. By contrast, a number of normalization techniques have been defined in the field of speaker verification that operate on the resulting log-likelihood scores. In this paper, we provide a theoretical motivation for likelihood normalization due to the so-called "hubness" phenomenon and we evaluate the benefit of several normalization techniques on ASR accuracy for the 2nd CHiME Challenge task. We show that symmetric normalization (S-norm) reduces the relative error rate by 43% alone and by 10% after feature and model compensation.

Index Terms: noise-robust speech recognition, concentration of distances, hubness

1. Introduction

While automatic speech recognition (ASR) is now very effective in clean conditions, reverberant and noisy conditions still pose a great challenge [1]. Robust ASR techniques operate by compensating for the impact of reverberation and noise in the features, in the acoustic models, or both [2–5].

Among the earliest feature transforms, cepstral mean and variance normalization [6] reduce the variability of the features by equalizing their moments across all environments. Feature-space maximum likelihood linear regression (fMLLR) [7] and linear discriminant analysis (LDA) [8] provide more flexibility by generatively or discriminatively adapting the transform to the environment. Such linear transforms greatly increase robustness to channel mismatch, but less so to reverberation and noise which result in nonlinear distortion of the features. Reverberation and noise can be reduced prior to feature extraction by means of signal-space speech enhancement or source separation techniques [9, 10]. Feature-space nonlinear transforms such as vector Taylor series [11], stereo-based piecewise linear compensation for environments (SPLICE) [12] and feature-space minimum phone error (fMPE) [13] have also showed to be effective [14].

As for the transformation of the acoustic models, maximum a posteriori (MAP) [15] and maximum likelihood linear regression (MLLR) [7, 16] have been widely used to adapt the mean and/or the covariance of Gaussian mixture model (GMM) observation densities to the environment. Multicondition training, i.e., directly training GMMs on reverberant and noisy data, can work even better when sufficient training data are available. The best results are often achieved by hybrid approaches applying model compensation after feature compensation on the training and the test data [12, 17, 18].

A common trait of the above techniques is that they operate on the features or on the GMM parameters themselves. By contrast, a number of normalization techniques have been defined in the field of speaker verification that operate on the set of acoustic scores, i.e., on the likelihoods of the acoustic models given the feature vectors [19]. These techniques are fundamentally different from the above in that the normalized scores are not restricted to be proper likelihoods anymore, i.e., they may each integrate to a different value other than 1. This extra flexibility increases the discriminativity of the decision thresholds.

In this paper, we provide a theoretical motivation for likelihood normalization due to the so-called "hubness" phenomenon [21] and we evaluate the benefit of several normalization techniques on ASR accuracy. Compared to speaker verification, this raises the challenge of balancing the normalized GMM scores and the HMM state transition scores. The resulting gain in noise robustness is evaluated on Track 1 of the 2nd CHiME Speech Separation and Recognition Challenge [20]. The paper is organized as follows. We provide more background information about hubness and speaker verification in Section 2. We generalize existing score normalization techniques and hubness measures to the context of ASR in Section 3. We report experimental results in Section 4 and we conclude in Section 5.

2. Motivation

Hubness has been described and explored as a fundamental issue for machine learning in high dimensional spaces [21]. Most machine learning problems involve computing the "distance"[1] between points in feature and/or model spaces. Hubness is the phenomenon by which certain points called "hubs" have a small distance to an exceptionally large number of points while certain points called "anti-hubs" are far from all other points. It is related to the concentration of distances [22], which impairs the contrast of distances in high dimensional spaces. As dimensionality increases, pairwise distances become almost identical to each other thus making it difficult to distinguish between the farthest and the closest point. This phenomenon has been mathematically studied for Euclidean spaces and other l^p norms [22, 23] and empirically observed for many other forms of "distances" including negative log-likelihoods between feature and model spaces [24].

Proofs concerning concentration of distances have been formulated for dimensionality approaching infinity. However, the dimension does not need to be very large for hubness to occur. In the finite case, some points are expected to be closer to the

[1] This term refers here to a pairwise dissimilarity score, which does not necessarily satisfy the mathematical definition of a distance.

center of the space and at the same time closer, on average, to all other points [21]. Such points closer to the center have a high probability of being hubs, i.e., of appearing in nearest neighbor lists of many other points. Points which are further away from the center have a high probability of being anti-hubs, i.e., points that never appear in any nearest neighbor list. Hubness has been reported for dimensions as low as on the order of 10 in practice [25].

This phenomenon is exacerbated by distortions of the features and/or the models, which further reduce the contrast of distances. Nevertheless, it is essential to understand that it is a separate issue and that it may still occur beyond a certain dimension even after feature and model compensation.

This behavior has a negative impact on many machine learning tasks including classification [21], nearest neighbor based recommendation [26], outlier detection [21,27] and clustering [28]. Closer to ASR, it has been shown that the "Doddington zoo" effect [29] in speaker verification is connected to the hubness phenomenon [24]. In this context, audio samples from "wolves" easily impersonate other speakers (are close to many speaker models), while persons which are difficult to recognize are referred to as "goats" (because they appear far from all other speaker models). "Wolves" and "goats" are therefore essentially hubs and anti-hubs.

A general solution to reduce hubness lies in the normalization of the distances [25]. Several normalization techniques have been proposed in the field of speaker verification that aim to equalize the mean and the variance of the acoustic log-likelihood scores (see [19] for a review). The mean is normalized in such a way that the decision threshold between target and impostor speakers is always fixed to the same score while the variance is normalized to unity. The two fundamental operations are zero normalization (Z-Norm) [30,31], which reduces the score variability across the target models, and test normalization (T-norm) [32], which reduces the score variability across the test utterances. Combinations of these two operations have also been proposed, including test-dependent zero-score normalization (TZ-norm) and zero-dependent test-score normalization (ZT-norm) [33, 34], which consist of applying them one after the other, and symmetric normalization (S-norm) [35], which consists of summing the two normalized scores together. These normalization techniques as well as the ones developed in the machine learning community [25] have been shown to reduce hubness and to improve speaker verification accuracy [24].

3. Likelihood normalization

We propose to generalize to ASR the score normalization techniques introduced in speaker verification. While speaker verification outputs a single score per utterance, ASR outputs a sequence of words, each of which is a sequence of acoustic units, e.g., phones. Each unit is modeled by a context-dependent hidden Markov model (HMM) with GMM observation densities. This raises the challenge of balancing the normalized GMM scores and the HMM state transition scores.

3.1. Formulation of ASR

We formulate ASR as the following weighted maximum a posteriori (MAP) decoding problem:

$$\arg\max_{\mathbf{w},\mathbf{s}} \sum_t Q(\mathbf{x}_t|\mathbf{s}_t) + \alpha \log P(\mathbf{s}|\mathbf{w}) + \alpha\beta \log P(\mathbf{w}) \quad (1)$$

with $\mathbf{x}_t$ the observed feature vector at time t, $\mathbf{s}$ the sequence of HMM states $(\mathbf{s}_t)$, and $\mathbf{w}$ the word sequence. We call $Q(\mathbf{x}_t|\mathbf{s}_t)$ the observation scores and $\log P(\mathbf{s}|\mathbf{w})$ the state transition score and $\log P(\mathbf{w})$ is the language score. We introduce the weight α to balance the observation scores and the other scores, while β is the language model weight as classically used in ASR [36]. Conventional decoding is achieved by setting $\alpha = 1$ and $Q(\mathbf{x}_t|\mathbf{s}_t) = \log P(\mathbf{x}_t|\mathbf{s}_t)$ where $P(\mathbf{x}_t|\mathbf{s}_t)$ is the GMM density associated with state $\mathbf{s}_t$.

3.2. Normalizing the observation scores

As we shall see, the features and the acoustic models used for robust ASR exhibit a certain amount of hubness. This translates into two separate issues.

Firstly, certain HMM states have systematically high (resp. low) observation scores even when they do not (resp. do) correspond to the true states behind the feature vectors, which is a major problem for ASR. The balance between the observation scores for different states can be restored via the Z-norm

$$Q^Z(\mathbf{x}_t|\mathbf{s}_t = i) = \frac{\log P(\mathbf{x}_t|\mathbf{s}_t = i) - \mu_i}{\sigma_i} \quad (2)$$

where μ_i and σ_i denote the mean and the standard deviation of $\log P(\mathbf{x}_t|\mathbf{s}_t = i)$ over "impostor" feature vectors $\mathbf{x}_t$, respectively. These quantities are estimated on the full development set and we define impostors in this context as all feature vectors whose corresponding state (according to forced alignment with the true word sequence) is not i. This normalization equalizes the decision threshold for all states.

Secondly, certain feature vectors result in more (resp. less) contrasted observation scores for all states. Although this is less of a problem than above, this may still affect the estimated word sequence to some extent by giving more (resp. less) importance to the observation scores in the MAP decoding rule (1). The balance between the observation scores and the other scores may be restored by the T-norm

$$Q^T(\mathbf{x}_t|\mathbf{s}_t = i) = \frac{\log P(\mathbf{x}_t|\mathbf{s}_t = i) - \mu_{it}}{\sigma_{it}} \quad (3)$$

where μ_{it} and σ_{it} denote the mean and the standard deviation of $\log P(\mathbf{x}_t|\mathbf{s}_t = i)$ over "impostor" states i, respectively. These quantities are computed on each time frame in the test data and we define impostors in this context as all states except i, that is $I - 1$ states where I is the total number of states of all HMMs.

Finally, both normalization techniques may be combined into the S-norm:

$$Q^S(\mathbf{x}_t|\mathbf{s}_t = i) = Q^Z(\mathbf{x}_t|\mathbf{s}_t = i) + Q^T(\mathbf{x}_t|\mathbf{s}_t = i). \quad (4)$$

The normalized observation scores are used in place of the original scores $Q(\mathbf{x}_t|\mathbf{s}_t)$ and MAP decoding is achieved as in (1).

We tried several variants of the above normalization techniques, such as computing the T-norm over all states instead of impostors only and computing the Z-norm over each test utterance or over all feature vectors in the development set instead of impostors only. We also tried the TZ-norm and the ZT-norm as alternatives to the S-norm. All these variants resulted in slightly lower ASR performance than the corresponding original norms above and they are not explored hereafter.

3.3. Hubness measures

In the following, we assess the impact of likelihood normalization not only on ASR accuracy but also on hubness. A measure was defined in [21] to quantify the strength of the hubness phenomenon in a given space. Since ASR involves two different spaces for features and models, we adapt this definition into two complementary measures drawn from the scores $Q(\mathbf{x}_t|\mathbf{s}_t = i)$ of all feature vectors and all states. We fix $k = 5$ as in [25].

For each feature vector $\mathbf{x}_t$, we first determine the k most likely states. For each state i, we then define the k-occurrence M_i as the number of feature vectors for which it is among the k most likely. Model-space hubness S_{model} is computed as the skewness of the distribution of M_i:

$$S_{\text{model}} = \frac{\mathbb{E}[(M_i - \mu_M)^3]}{\sigma_M^3} \tag{5}$$

with μ_M its mean and σ_M its standard deviation. Values close to zero indicate low hubness, while large (positive or negative) values indicate high hubness.

Similarly, we find the k most likely feature vectors for each state i and we count the k-occurrence N_t as the number of states for which $\mathbf{x}_t$ occurs among the k most likely feature vectors. Feature-space hubness S_{feature} is obtained as:

$$S_{\text{feature}} = \frac{\mathbb{E}[(N_t - \mu_N)^3]}{\sigma_N^3}. \tag{6}$$

4. Experimental evaluation

We assess the impact of likelihood normalization on noise robustness. Evaluation is conducted on Track 1 of the 2nd CHiME Speech Separation and Recognition Challenge [20].

4.1. Data and task

The Challenge aims to recognize distant speech in a real home with reverberation and multisource background noise. The target utterances are 6-word commands of the form <command> <color> <preposition> <letter> <digit> <adverb> read by 34 speakers at 6 different signal-to-noise ratios (SNRs) from -6 to +9 dB. The speaker is assumed to be known and the task is to report the letter and digit keywords, which are the most easily confusable words. Accuracy is measured as the percentage of correctly recognized keywords. Three training sets are provided, each involving 500 utterances per speaker: a clean training set, a reverberated (noiseless) training set, and a noisy (reverberant) training set. The development set and the test set each involve 600 reverberated noisy utterances per SNR.

4.2. Feature and model compensation

The proposed normalization techniques are evaluated alone and in combination with feature and model compensation. To evaluate the effect on feature compensation, we consider both the original noisy data and enhanced data. Speech enhancement is applied to the development and test datasets and to the noisy training set using the Flexible Audio Source Separation Toolbox (FASST) [37] with the same settings as in [18]. Uncertainty propagation is employed to robustly extract feature vectors from the separated signals as detailed in [18].

Three sets of acoustic models are considered: clean models trained on the clean training set, reverberated models (abbreviated “reverb”) trained on the reverberated training set, and multicondition models (“multi”) trained on the original/enhanced noisy training set (all SNRs included). Multicondition training was indeed found to perform better than model adaptation techniques for this Track, due to the large amount of training data [20]. All models are speaker-dependent and SNR-independent.

4.3. Algorithm settings

The normalization factors for the Z-norm (and consequently for the S-norm) are assumed to be speaker-dependent. In order to evaluate the impact of data mismatch on the estimation of these factors, they are trained either in an SNR-dependent fashion on the development data for the same speaker and SNR as the test utterance or in an SNR-independent fashion on all development data of the same speaker.

Decoding is performed using the baseline HTK setup provided by the Challenge organizers [20]. The features are 39-dimensional vectors consisting of 12 Mel-frequency cepstral coefficients (MFCCs), log-energy, delta and acceleration coefficients. The acoustic models are word-level left-to-right HMMs with 2 states per phoneme. Each state is modeled by a GMM with 7 Gaussians with diagonal covariance.

Due to the constrained syntax, the language model weight has no impact and it is fixed to $\beta = 1$. The weight α is assumed to be SNR-independent and it is optimized on development data in the range of [0,50] with a step of 1 for the original scores and in the range of [0,4] with a step of 0.2 for the normalized likelihoods[2]. The optimal weight is found to vary between 0 and 11 in the former case and between 0 and 1.6 in the latter case depending on the data (noisy or enhanced), on the model, and on the normalization technique. The weight found on development data is subsequently applied to the test data.

4.4. ASR results

The average ASR performance achieved for different data, models, and normalization techniques is shown in Table 1, with the best (statistically significant) results in each column highlighted in bold. The baseline test set accuracy ranges from 17.64% without compensation (noisy data, clean model) to 85.85% with feature and model compensation (enhanced data, multicondition model).

SNR-dependent and SNR-independent normalization are seen to perform very similarly. We hence focus on SNR-independent results hereafter.

Applied alone, the change of the value of α and the T-norm have no significant effect on accuracy. The Z-norm improves performance on all data in combination with clean or reverberated models, but not with multicondition models.

The S-norm is the only norm that improves results on all data and for all models. Compared to the baseline, it achieves a test set accuracy of 52.74% without compensation and 87.25% with feature and model compensation, that is a relative error rate reduction of 43% and 10%, respectively. This is in agreement with previous work in the field of machine learning which has shown that it is decisive to normalize distances with respect to both arguments [25].

The latter result is analyzed as a function of the SNR in Table 2, with the best (statistically significant) results in each column highlighted in bold. Likelihood normalization has no significant effect at high SNRs but it increases performance up to 19% relative at low SNRs, that is twice as much as on average over all SNRs.

4.5. Analysis of hubness

The measured hubness is reported in Table 3. The original values are on the order of 2 to 4, which is moderate to large [25]. For a given dataset and model, the lower the model hubness

[2] $\alpha = 0$ means that the left-to-right state order and the word order are enforced, but state durations are not constrained anymore.

Dataset		Noisy development set			Noisy test set			Enhanced development set			Enhanced test set		
Acoustic model		Clean	Reverb	Multi	Clean	Reverb	Multi	Clean	Reverb	Multi	Clean	Reverb	Multi
No norm	$\alpha = 1$	17.72	57.42	69.19	17.64	58.53	69.47	26.10	77.15	85.17	25.93	79.21	85.85
	α opt.	18.53	57.97	69.28	18.72	58.33	69.42	26.51	77.51	85.24	27.03	79.14	85.82
Z-norm	SNR-dep.	**58.92**	70.15	66.76	**57.67**	68.76	65.94	**66.49**	78.38	81.25	**66.44**	78.11	81.26
	SNR-ind.	**58.06**	69.63	66.22	**57.67**	68.57	66.07	**66.31**	78.38	81.00	**66.31**	78.25	81.46
T-norm		18.54	56.86	69.01	18.74	57.74	69.24	26.15	76.68	84.47	26.32	78.68	85.26
S-norm	SNR-dep.	54.39	**71.39**	**73.72**	53.04	**71.14**	**73.46**	64.61	**83.08**	**86.65**	64.35	**83.78**	**87.26**
	SNR-ind.	53.67	**71.35**	**73.64**	52.74	**71.13**	**73.28**	64.75	**83.18**	**86.60**	64.71	**83.68**	**87.25**

Table 1: Keyword accuracy (in %) for all data and models averaged over all SNRs.

Multicondition	Enhanced development set						Enhanced test set					
acoustic model	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
No norm	73.92	**79.17**	84.33	**89.58**	**91.12**	**93.33**	75.75	77.75	**85.33**	**90.50**	**92.50**	**93.08**
S-norm (SNR-ind.)	**78.33**	**81.25**	**86.33**	**89.25**	**91.50**	**92.92**	**79.58**	**82.08**	**87.17**	**90.42**	**91.67**	**92.58**

Table 2: Keyword accuracy (in %) for the enhanced data and the multicondition acoustic models as a function of the SNR.

Measure	Model hubness S_{model}						Feature hubness S_{feature}					
Dataset	Noisy test set			Enhanced test set			Noisy test set			Enhanced test set		
Acoustic model	Clean	Reverb	Multi	Clean	Reverb	Multi	Clean	Reverb	Multi	Clean	Reverb	Multi
No norm	4.31	2.69	1.67	3.95	2.06	1.53	2.03	2.65	4.00	1.93	2.27	2.67
Z-norm (SNR-ind.)	**1.67**	1.79	1.85	**1.59**	1.67	1.68	2.03	2.65	4.00	1.93	2.27	2.67
T-norm (SNR-ind.)	4.31	2.69	1.67	3.95	2.06	1.53	1.10	0.90	0.61	1.06	0.75	0.60
S-norm (SNR-ind.)	1.81	**1.67**	**1.53**	1.73	**1.49**	**1.39**	1.54	1.22	1.81	1.46	1.05	1.02

Table 3: Test set hubness for all data and models.

the better the ASR performance. Feature compensation, model compensation, and S-normalization all reduce model hubness. The smallest model hubness $S_{\text{model}} = 1.39$ is achieved when jointly using these three techniques. This is confirmed in Figure 1 (horizontal axis) which depicts the measured hubness before and after normalization for each utterance in the enhanced test set. Feature hubness, on the other hand, appears to be unrelated to ASR performance. This confirms that biases in the observation scores are more crucial to address than imbalance with the other scores.

5. Conclusion

We explored the use of likelihood normalization techniques for speaker verification in the context of robust ASR. We adapted these techniques to the observation scores in ASR and we balanced the other terms in the MAP decoding rule via a weight α. We showed that joint normalization of the scores across both the models and the features significantly improves performance even after feature and model compensation. ASR performance appears to be well correlated with model hubness, with lower values of model hubness resulting in greater ASR accuracy. Future work will focus on computing the S-norm from a subset of states for greater computational efficiency [25], on evaluating it on larger-vocabulary tasks, and on finding improved normalization techniques to further reduce model hubness.

6. Acknowledgment

This work was supported by FWF grant P24095 and by the LangTERRA project (FP7-REGPOT-2011-1). We thank D. T. Tran for the enhanced features and D. Jouvet for his advice.

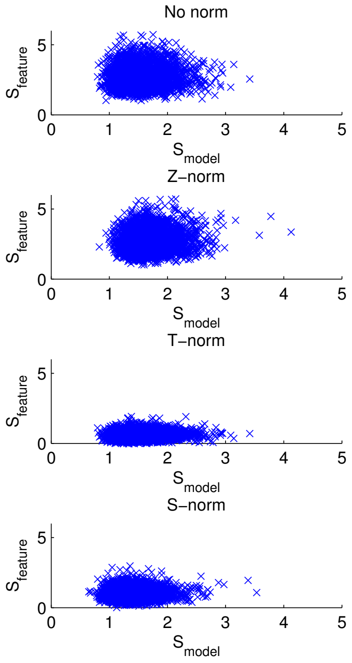

Figure 1: Model and feature hubness measured on the enhanced test set with multicondition acoustic models. Each dot corresponds to one utterance.

7. References

[1] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding, part 1," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 75–80, May 2009.

[2] J. Droppo and A. Acero, "Environmental robustness," in *Handbook of Speech Processing*. Springer, 2007, pp. 653–680.

[3] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Wiley, 2009.

[4] L. Deng, "Front-end, back-end, and hybrid techniques for noise-robust speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*. Springer, 2011, pp. 67–99.

[5] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.

[6] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1–3, pp. 133–147, 1998.

[7] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[8] R. Häb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992, pp. 13–16.

[9] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[10] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind speech separation*. Springer, 2007.

[11] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 1996, pp. 733–736.

[12] L. Deng, A. Acero, M. Plumpe, and X. D. Huang, "Large vocabulary speech recognition under adverse acoustic environments," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2000, pp. 806–809.

[13] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 961–964.

[14] Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME Challenge benchmark," in *Proc. 2nd Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, 2013, pp. 19–24.

[15] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[16] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[17] R. Astudillo and D. Kolossa, "Uncertainty propagation," in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*. Springer, 2011, pp. 35–62.

[18] D. T. Tran, E. Vincent, and D. Jouvet, "Extension of uncertainty propagation to dynamic MFCCs for noise-robust ASR," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[19] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.

[20] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: An overview of challenge systems and outcomes," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013, pp. 162–167.

[21] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *Journal of Machine Learning Research*, vol. 11, pp. 2487–2531, 2010.

[22] D. François, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, pp. 873–886, 2007.

[23] C. Aggarwal, A. Hinneburg, and D. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Proc. 8th Int. Conf. on Database Theory (ICDT)*, 2001, pp. 420–434.

[24] D. Schnitzer, A. Flexer, and J. Schlüter, "The relation of hubs to the Doddington zoo in speaker verification," in *Proc. 21st European Signal Processing Conf. (EUSIPCO)*, 2013, pp. 1–5.

[25] D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer, "Local and global scaling reduce hubs in space," *Journal of Machine Learning Research*, vol. 13, pp. 2813–2844, 2012.

[26] A. Flexer, D. Schnitzer, and J. Schlüter, "A MIREX meta-analysis of hubness in audio music similarity." in *Proc. 13th Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2012, pp. 175–180.

[27] A. Flexer and D. Schnitzer, "Using mutual proximity for novelty detection in audio music similarity," in *Proc. 6th Int. Workshop on Machine Learning and Music (MML)*, 2013, pp. 31–34.

[28] N. Tomašev, M. Radovanović, D. Mladenić, and M. Ivanović, "The role of hubness in clustering high-dimensional data," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2013, pp. 183–195.

[29] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," DTIC Document, Tech. Rep., 1998.

[30] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1988, pp. 595–598.

[31] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[32] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, 2000.

[33] R. J. Vogt, B. J. Baker, and S. Sridharan, "Modeling session variability in text-independent speaker verification," in *Proc. Interspeech*, 2005, pp. 3117–3120.

[34] R. Zheng, S. Zhang, and B. Xu, "A comparative study of feature and score normalization for speaker verification," in *Proc. Int. Conf. on Biometrics (ICB)*, 2006, pp. 531–538.

[35] S. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," in *Proc. Odyssey*, 2010, pp. 76–82.

[36] L. R. Bahl, R. Bakis, F. Jelinek, and R. L. Mercer, "Language-model/acoustic channel balance mechanism," *IBM Technical Disclosure Bulletin*, vol. 23, no. 7B, pp. 3464–3465, 1980.

[37] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.