A Study of Likelihood Ratio Calibration in High Vocal Effort Speech for a Modern Automatic Speaker Recognition System

Miranti Indar Mandasari, Rahim Saeidi, David A. van Leeuwen
Centre for Language and Speech Technology, Radboud University Nijmegen, the Netherlands
{m.mandasari|r.saeidi|d.vanleeuwen}@let.ru.nl

The production of speech is not only influenced by various intrinsic factors such as semantics, dialect, human perspective and emotion, but also by extrinsic factors such as environmental conditions and transmission channel. In certain acoustic conditions, the vocal effort of a speaker tends to be raised in order to overcome environmental hindrances such as a presence of noise or a long distance between the speaker and listener. There have only been a few studies on speaker recognition under non-neutral speech production conditions (i.e., high or low vocal effort and speech under stress) (Hansen, 2011). However, in real forensic cases, it can occur that the incriminating recording is made with high vocal effort, which then has to be dealt with in speaker comparison.

This paper presents a study of the effect of high vocal effort speech to the automatic speaker recognition system performance, considering the likelihood ratio (LR) calibration aspect. Using the most recent algorithm in the field (Burget, 2011 and Dehak, 2011), the calibration performance of the system is evaluated on both high and normal vocal effort conditions of the latest NIST speaker recognition evaluation (SRE) (http://www.nist.gov/itl/iad/mig/sre.cfm).

State-of-the-art automatic speaker recognition system

In this paper, we use an automatic speaker recognition system based on i-vector framework (Dehak, 2011) using auditory induced acoustic features, and probabilistic linear discriminant analysis (PLDA) modeling with a similar setup as we used in our latest work (Mandasari, 2012). An i-vector is a representation of a speech utterance in a relatively low-dimensional space called total variability space which consists of both speaker and channel variability. It was firstly introduced by Dehak (2009) and has become a mainstream in speaker recognition field since then. PLDA modeling is a probabilistic approach to model the i-vector distribution in the form of a multivariate Gaussian, and our implementation follows the approach of Burget (2011). The PLDA model produces log likelihood scores which we treat as un-calibrated scores that can be calibrated by a linear transformation (Brümmer, 2006).

Log likelihood ratio calibration

Rodriguez (2007) argues that a set of scores produced by an automatic speaker recognition system should be calibrated in order to produce more reliable and less misleading LRs. Even though the output scores from PLDA system are formulated as a log likelihood ratio, a calibration is still necessary in order to be used in, e.g., forensic speaker comparison.

Experiment setup and results

We measure the cost of scores calibration by using the *log likelihood ratio calibration cost* (Cllr) metric (Brümmer, 2006; for an introduction see Van Leeuwen, 2007). The calibration performance of the system is evaluated in terms of miscalibration cost, which is the difference between the actual Cllr and the minimum Cllr that can be obtained by optimal transformation of the evaluated log LR values. We used two disjoint databases for training the calibration and evaluation. For training the calibration parameters, we used 'det 7' core condition of NIST SRE 2008 corpus which includes normal vocal effort utterances in the model and test sides. These calibration parameters are then ap-

plied to the evaluation data that comes from NIST SRE 2010 extended trials for 'det 5' and 'det 6' conditions which contain utterances in normal and high vocal effort in the test side, respectively.

Table 1 presents the calibration performance of the automatic speaker recognition system. Calibration was investigated in two different conditions, matched and mismatched. Since the calibration parameters were trained on normal vocal effort speech in model and test, the *matched* calibration condition was obtained by evaluating the calibration performance on NIST SRE 2010 'det 5' condition (normal vocal effort in model and test sides). Evaluating the calibration parameters on 'det6' condition of NIST SRE 2010 is considered as *mismatched* calibration condition. In this 'det 6' condition, the model side is trained with normal vocal effort speech while the test side comes from high vocal effort speech.

Table 1. Calibration performance of the i-vector and PLDA based speaker recognition system in matched and mismatched conditions evaluated on NIST SRE 2010 'det 5' and 'det 6', respectively.

Gender	*Calibration condition*	*Vocal effort (model-test)*		*Minimum Cllr*	*Actual Cllr*	*Miscalibration cost*
		Calibration	*Evaluation*			
Male	Matched	normal-normal	normal-normal	0.0703	0.0768	0.0065
	Mismatched		normal-high	0.1342	0.2067	0.0724
Female	Matched	normal-normal	normal-normal	0.1216	0.1332	0.0117
	Mismatched		normal-high	0.1881	0.2592	0.0711

Compared to the matched condition, the minimum Cllr for both genders are found to be increased by 0.06 in the mismatched vocal effort condition, which is about 100% and 50% higher for male and female cases, respectively. This result shows that the discrimination performance of an automatic speaker recognition system is largely affected by the high vocal effort speech. The results on the actual Cllr values in the mismatched condition demonstrate that the calibration is affected by the high vocal effort as well, with an addition of 0.07 from the matched condition in both genders. These experiment results show that the calibration problem for high vocal effort conditions is as difficult as that of discrimination, thus motivating further research to find more efficient techniques to overcome the mismatch in vocal effort.

References

Burget, L., O. Plchot, S. Cumani, O. Glembek, P. Matejka and N. Brümmer. (2011). Discriminatively trained probabilistic linear discriminant analysis for speaker verification. In proc. of ICASSP '11, p. 4832-4835.

Brümmer, N. and J. du Preez. (2006). Application-independent evaluation of speaker detection, Computer, Speech and Language, vol 20, p. 230-275.

Dehak, N., R. Dehak, P. Kenny, N. Brümmer, P. Ouellet and P. Dumouchel. (2009). Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In proc. of the 10th Annual Conference of the International Speech Communication Association, p. 1559-1562.

Dehak, N., P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet. (2011). Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, p. 788-798.

Hansen, J.H.L., A. Sangwan and W. Kim. (2011). Speech Under Stress and Lombard Effect: Impact and Solutions for Forensic Speaker Recognition. Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism, chapter 5, p. 103: Springer Verlag.

Van Leeuwen, D. and N. Brümmer. (2007). An introduction to application-independent evaluation of speaker recognition systems. Speaker Classification I, p. 330-353: Springer.

Mandasari, M. I., M. McLaren and D. A. van Leeuwen. (2012). The Effect of Noise on Modern Automatic Speaker Recognition Systems. In proceedings of ICASSP '12, p. 4249-4252.

Rodriguez, J. G. and D. Ramos. (2007). Forensic automatic speaker classification in the "coming paradigm shift". Speaker Classification, p. 205-217: Springer.

Traunmüller, H. and A. Eriksson. (2000). Acoustic effects of variation in vocal effort by men, women, and children. The Journal of the Acoustical Society of America, vol. 107, p. 3438.