ASSESSMENT OF SPEAKER VERIFICATION DEGRADATION DUE TO PACKET LOSS IN THE CONTEXT OF WIRELESS MOBILE DEVICES

Nicholas W. D. Evans[1], *John S. Mason*[1], *Roland Auckenthaler*[2] *and Robert Stapert*[3]

[1]School of Engineering, University of Wales Swansea, UK
{eeevansn, j.s.d.mason}@swansea.ac.uk, http://eegalilee.swansea.ac.uk
[2]Ubiquity, Newport, UK
rauckenthaler@hotmail.com
[3]Aculab, Milton Keynes, UK
robert.stapert@aculab.com

ABSTRACT

This paper considers the adverse effects on speaker verification accuracy caused by two independent forms of speech signal degradation common in mobile communications. The two forms are packet loss in the communications system and ambient noise at the wireless device. The effects of these degradations are assessed independently on a common database of 2000 speakers. Baseline verification performances in terms of equal error rates (EER) show negligible degradation until over 75% of test feature vectors are lost. The EER grows from 3% to just 5% when the loss reaches 88%. In contrast, adding a relatively small amount of noise to the test speech (15dB SNR), with otherwise identical experimental conditions, results in a rise in the EER to 36%. In this latter case, simple speech enhancement leads to a reduction in EER to 21%. The main conclusion of this work is that, for speech-based verification, typical packet loss is likely to incur a negligible degradation in accuracy when compared with the degradation that is associated with typical ambient noise conditions.

1. INTRODUCTION

Biometric measures are a potentially useful and reliable approach to person verification. Through the ever-increasing use of Internet-enabled wireless mobile devices, everyday tasks such as banking and shopping are conducted without face-to-face or personal contact. It is then more difficult to verify a persons identity using conventional methods, such as signatures or photo identity cards. Over IP networks, both speech and visual-based biometrics are viable alternative approaches to verification. This paper focuses on speech biometrics. In future, speaker verification systems may be used to provide an additional layer of security for on-line commercial applications or for eaves dropping of real-time conversations. These two examples illustrate the wide domain in which speech systems operate. In the context of wireless mobile devices, systems are inherently susceptible to wide variations in transmission conditions.

Most wireless mobile networks are susceptible to packet loss to some degree. Whilst there exist many strategies to combat packet loss, such as re-transmission or packet recovery [1, 2, 3], on-line identity verification applications may still operate effectively from semi real-time voice streams. This is possible because there is no intrinsic requirement on latency in the case of re-transmission, and there is evidence that speaker verification systems are resilient to packet loss [4]. Significant data loss inevitably has a detrimental impact and in this paper speaker verification accuracy is assessed against the level of packet loss.

The packet loss scenario is then contrasted with degradation coming from additive noise. The degrading effect of ambient noise on automatic speech and speaker recognition is widely acknowledged and known to be large even for relatively low noise levels. Thus a comparison is made between the two forms of degradation by using otherwise identical experimental conditions.

The remainder of this paper is organised as follows. Section 2 addresses the effect of packet loss in typical wireless and IP networks on speaker verification. Section 3 addresses additive noise and speech enhancement. Experimental work on the 2000 speaker SpeechDat Welsh [5] database is presented in Section 4 with results of experiments with both simulated packet loss and speech enhancement after contamination by additive real car noise. Conclusions are presented in Section 5.

2. PACKET LOSS

Some degree of packet loss is inherent in mobile networks. Lost packets might be caused by variable transmission conditions, or the hand-over between neighbouring cells as a

wireless mobile device roams about the network.

Approaches dealing with packet loss recovery are generally controlled by the routing protocol adopted in the network architecture. For automatic speech recognition applications where time sequence information is more critical, packet loss might have a significant impact on performance. Lost packets might then be re-transmitted or some form of compensation employed [1, 2, 3]. In contrast, for speaker verification a limited degree of packet loss might not have a too detrimental effect, particularly in text-independent mode. This form of speaker verification is generally less dependent on time sequence information and there is some evidence in a related study of computational efficiency [4], that speaker verification systems might be relatively insensitive to packet loss. One potential anomaly in this hypothesis, equally applicable to both speech and speaker recognition, is the effect of lost packets on dynamic features which are dependent on their static counterparts over some small window, typically in the order of 100ms or more. Unless appropriately compensated, corrupt dynamic features can lead to performance degradation. The work presented in this paper addresses the problem of packet loss without recovery in a speaker verification context. As in [4], speech feature vectors are discarded rather than raw speech data frames. These features are assumed to be calculated in the wireless mobile device and tailored to automatic recognition systems similar to the ETSI-AURORA standard [6].

In the loss simulation features are discarded with varying temporal resolutions to simulate packet loss. Experiments are performed with a conventional implementation of a Gaussian mixture model (GMM) [7] as used by most of today's text-independent speaker verification systems.

3. ADDITIVE NOISE

The second degradation considered here typifies the conditions under which wireless mobile devices are commonly used, namely with a meaningful level of background noise. The consequences of such additive noise are:

- direct contamination of the speech signal, and
- induced changes in the speaking style of the persons subjected to the noise, known as the Lombard reflex [8].

In these experiments noise is added to the speech recordings thereby minimising any Lombard effects. The noise is added at a moderate level of 15dB SNR. Subsequently, for completeness, a simple speech enhancement process is applied to the degraded signal.

The form of enhancement considered here has the option of returning the speech to the time domain. Such an approach might lead to sub-optimal compensation in terms of recognition performance but none the less offers benefits in terms of integration into existing systems and communications networks.

Perhaps the first notable work in this field is that of Boll [9] and Berouti *et al* [10] both in 1979. Speech enhancement for human-to-human conversation was performed by an approach still known today as spectral subtraction. Subsequently, Lockwood and Boudy [11] applied spectral subtraction extensively to automatic speech recognition.

There are many approaches and applications of spectral subtraction. Of particular interest here is an implementation of spectral subtraction termed quantile-based noise estimation (QBNE), proposed by Stahl *et al* [12]. QBNE is an extension of the histogram approach presented by Hirsch and Ehrlicher [13]. The main advantage of these approaches is that an explicit speech, non-speech detector is not required. Noise estimates are continually updated during both non-speech *and* speech periods from frequency-dependent, temporal statistics of the degraded speech signal. An efficient implementation of QBNE, important in the context of mobile systems, is described in [14].

4. EXPERIMENTAL RESULTS

4.1. Database

The experimental work presented in this paper was performed on the SpeechDat Welsh database [5]. The data consists of 2000 speakers recorded over a fixed telephony network. One thousand of the 2000 speakers were used to create a world model and the other 1000 speakers used for speaker model training and testing. Training was performed on approximately 30 seconds of phonetically rich sentences per speaker with a total of about 8 hours for the world model. Two separate text independent tests used either:

- a 4-digit string, or
- a single digit

per speaker per test, giving 1000 tests per experiment. Features are standard MFCC-14 static concatenated with 14 dynamic coefficients.

4.2. Packet Loss and Additive Noise Degradations

To simulate packet loss, a varying percentage of speech feature vectors are discarded from the test set. No attempt is made to recover these lost vectors although the minimum number of feature vectors per test is capped to two. Some results are presented in Figure 1. The detection error tradeoff (DET) curves show the system to be highly resilient with minimal increases in error rates until over 75% of the feature vectors are lost, the first three profiles being very close together. This is true for both plots: (a) the longer 4 digit

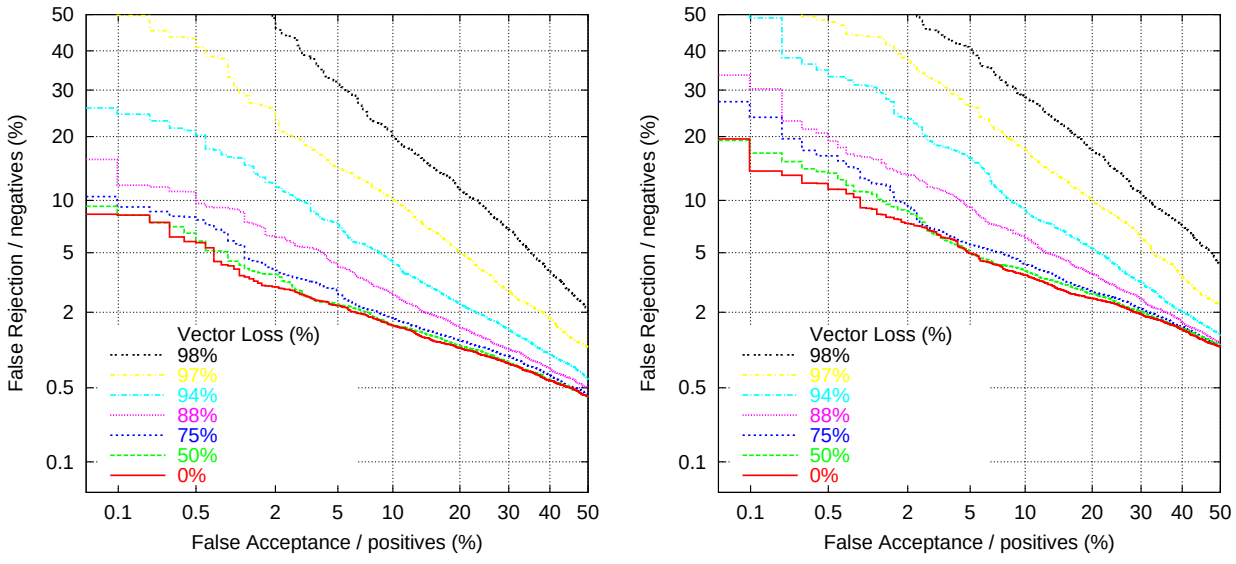

Figure 1: Speaker verification performance for varying degrees of feature vector loss, 0 up to 98% (with a minimum of 2 feature vectors maintained in all tests) for (a) 4 digit string tests (b) single digit tests.

string test utterances and (b) the shorter, single digit test utterance. Interestingly, in both cases, the profiles diverge toward the left. Considering the 4 digit case (left plot), this indicates that for operating points accepting high false acceptances in return for lower false rejections, the system is particularly robust against packet loss: just 2% false rejections with 50% false acceptances at the extreme case of 98% data loss. Evidence is presented again in Figure 2 where the equal error rates are plotted against percentage vector loss and it is clear that performance begins to degrade only after over 75% of the vectors are lost. This is very much in line with the findings of McLaughlin *et al* [4] who report that a factor of 20 loss can be tolerated before meaningful speaker verification degradation occurs.

To simulate speaker verification in adverse conditions, the test data were artificially contaminated with car noise at a moderate level of approximately 15dB SNR. Figure 3 illustrates the effects. The three profiles are for the original telephony test data (bottom profile) the contaminated test data (top profile) and the contaminated data after processing with the speech enhancement approach outlined above (middle profile). Clearly the levels of performance degradations are marked, even after compensation. This serves to illustrate how relatively small the degradation from packet loss might prove to be, in relation to additive noise.

5. CONCLUSIONS

This paper considers the adverse effects on speaker verification accuracy due to two independent forms of signal degradation common in mobile communications. The two forms are packet loss in the communications system and ambient noise at the wireless device. From previous work it could be deduced that ambient noise is likely to be far more detrimental than packet loss for most practical scenarios. This has indeed been supported by this work: a direct comparison of various packet-loss situations with a moderate ambient noise situation.

Using an otherwise identical experimental set-up the degradation due to packet loss proves to be negligible compared to the loss coming from a test to training mismatch of 15dB SNR. Even in the unlikely scenario of 98% feature vector loss (capped to a minimum of two frames per test utterance) the speaker verification performance using a 4 connected digit test degrades to just 15% (19% for the single digit test token). This is to be compared with the 36% (40% for single digit testing) in the case of additive noise in the region of just 15dB SNR. A simple non-linear spectral subtraction process improves the latter results to 21% (32%).

It is clear that even without any form of recovery, packet loss is unlikely to be significant in speaker verification performance when compared directly to the adverse effects of additive noise.

6. REFERENCES

[1] C. Perkins, O. Hodson, and V. Hardman, "A Survey of Packet Loss Recovery Techniques for Streaming Audio," *IEEE Network Magazine*, vol. 12, no. 5, pp. 40–48, 1998.

[2] B. Milner and S. Semnani, "Robust Speech Recognition over IP Networks," in *Proc. ICASSP*, 2000.

[3] P. Mayorga, R. Lamy, and L. Besacier, "Recovering of Packet Loss for Distributed Speech Recognition," in *Proc. EUSIPCO*, 2002.

[4] J. McLaughlin, D. A. Reynolds, and T. Gleason, "A Study of Computation Speed-ups of the GMM-UBM Speaker Recognition System," in *Proc. Eurospeech*, 1999, vol. 3, pp. 1215–1218.

[5] R. J. Jones, J. S. D. Mason, R. O. Jones, L. Helliker, and M. Pawlewski, "SpeechDat Cymru: A large-scale Welsh telephony database," in *Proc. LREC Workshop: Language Resources for European Minority Languages*, 1998.

[6] D. Pearce, "Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI standards activities for Distributed Speech Recognition Front-ends," in *Applied Voice Input/Output Society Conference*, 2000.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing, Academic Press*, vol. 10(1-3), pp. 19–41, 2000.

[8] J-C. Junqua, "The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers," *J. Acoust. Soc. Am.*, vol. 93, pp. 510–524, 1993.

[9] S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," *IEEE Trans. on ASSP*, vol. 27(2), pp. 113–120, 1979.

[10] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," in *Proc. ICASSP*, 1979, pp. 208–211.

[11] P. Lockwood and J. Boudy, "Experiments with a Non-linear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars," in *Proc. Eurospeech*, 1991, vol. 1, pp. 79–82.

[12] V. Stahl, A. Fischer, and R. Bippus, "Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering," in *Proc. ICASSP*, 2000, vol. 3, pp. 1875–1878.

[13] H. G. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition," in *Proc. ICASSP*, 1995, vol. 1, pp. 153–156.

[14] N. W. D. Evans, J. S. Mason, and B. Fauve, "Efficient Real-time Noise Estimation Without Explicit Speech, Non-speech Detection: An Assessment on the AURORA Corpus," in *Proc. Int. Conf. DSP*, 2002, vol. 2, pp. 985–988.

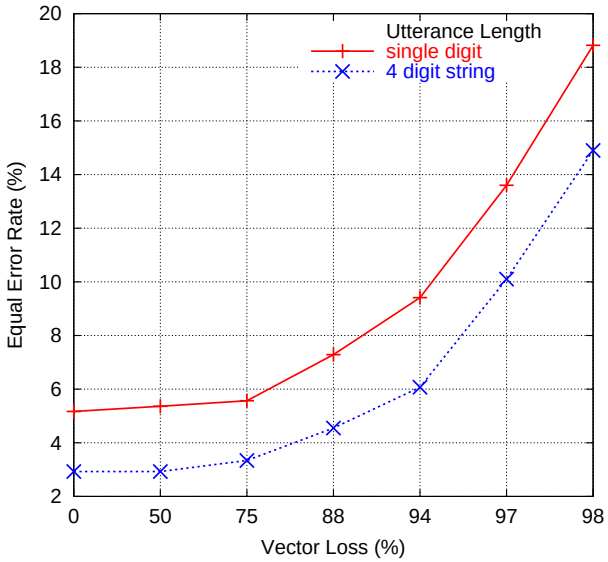

Figure 2: Equal error rate against feature vector loss (%) for test utterances of: 4 digit string (lower profile) and single digit utterance (upper profile). In all cases minimum test length maintained at two vectors.

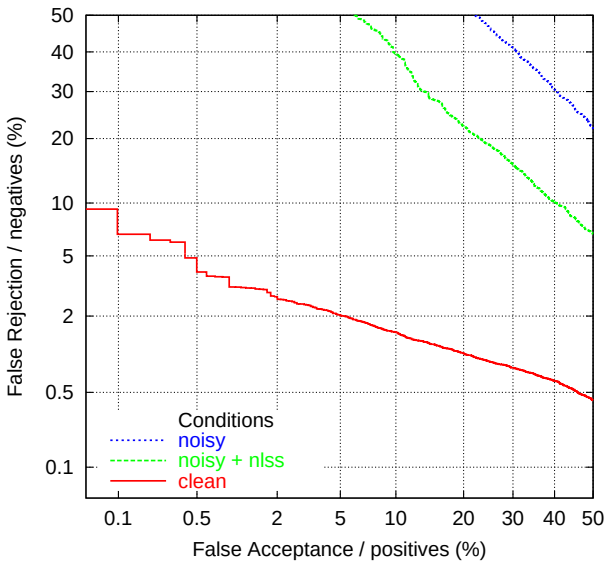

Figure 3: Speaker verification performance for the 4 digit string test set with top profile: 15dB SNR added noise, middle profile: 15dB SNR added noise plus speech enhancement, and bottom profile: original baseline.