A study of throat microphone performance in automatic speech recognition on motorcycles

T. Winkler, S. Pronkine, R. Bardeli, J. Köhler

Fraunhofer IAIS, Germany, Email: thomas.winkler@iais.fraunhofer.de

Introduction

In silent environments, automatic speech recognition usually provides a sufficient accuracy for different applications. Environmental noise, however, can decrease the performance of automatic speech recognition. Several algorithmic approaches exist to make speech recognition more robust to environmental noise (e.g., [2, 3, 5]). Still, a more direct way of increasing the robustness in loud environments can be the application of robust speech capturing components such as throat microphones.

On a motorcycle the speech signal can be distorted significantly by environmental noise. In the project MoveOn, which is co-funded by the European Commission, a multimodal system is designed and developed to provide communications and device control services to police motorcyclists on the move. Special attention has to be paid to the fact that most of the interaction with such a system has to be hands-free and that distraction has to be kept minimal. Speech, as hands-free input modality, is the central modality of the system. Thus, the reliability of the speech modality is of great importance even for high noise levels during acceleration periods and high speed levels. Therefore, noise robust transducer concepts are of general interest for the task in MoveOn.

In this paper, a high quality lavalier microphone and a noise robust throat microphone are evaluated and compared for clean and noisy environments. The data for this evaluation — a subset of the MoveOn database — was synchronously recorded for both types of microphones and reflects the difficult environmental conditions on the motorcycle.

Evaluation Data

The speech data for evaluation was recorded synchronously with a close-talk and a throat microphone. Both microphones and the database design are briefly described in this section.

Close-Talk Microphone

Standard close-talk microphones are well known to most people. Different transducer concepts as well as different characteristics, considering polar pattern, frequency response and dynamic range, exist. But the fundamental concept is to record airborne sound. In this work a close-talk microphone (AKG C417 - condenser lavalier microphone) with an almost linear frequency response in the spectral range of speech (up to about 8 kHz) is used. The microphone is capable of recording even high sound pressure levels with low distortion. The polar pattern is omnidirectional.

Throat Microphone

Throat microphones are more robust to environmental noise than usual close-talk microphones and, hence, are often used for communication in noisy environments in military and other applications. A throat microphone is put around the neck with the transducer placed on the larynx with slight pressure. Instead of airborne sound solid-born sound is picked up directly from the larynx. Thus, a throat microphone is less prone to environmental noise, but also lacks some specific frequency characteristics of speech. A standard single transducer throat microphone — the Alan AE 38 Throat Microphone — provides an alternative speech signal for evaluation.

Evaluation Database

The evaluation data is a subset of the MoveOn database described in [7]. This database was developed in the scope of the project MoveOn and consists of two main parts: 10 office recordings recorded in a silent environment and 39 recordings recorded on the motorcycle in a noisy environment. Three microphones synchronously recorded the speech, two close-talk microphones placed left and right of the speakers mouth within the motorcycle helmet and a throat microphone put around the neck. The recorded speech utterances are focused on but not limited to the domain and task in MoveOn. Airwind noise and noise from the engine of the motorcycle are dominant background noises in the database. But other typical noises like horns or traffic noise are also present. In this paper, the signals of the right close-talk microphone and the throat microphone are compared. Only sessions which contain both microphone channels are used in the evaluation.

Analysis of the Speech Signals

Close-talk and throat microphones have substantially different qualities regarding spectral characteristics and signal-to-noise ratio (SNR). The SNRs and some spectral characteristics of the microphones are compared in this section.

Signal-to-Noise Ratio

In Figure 1 the average SNR of the throat microphone is plotted versus the related SNR of the close-talk microphone for synchronously recorded utterances. The SNR is estimated by the NIST STNR, which is part of the NIST SPQA package [1]. For low SNRs the throat microphone is generally less affected by environmental noise, i.e., the average SNR is higher than the SNR for

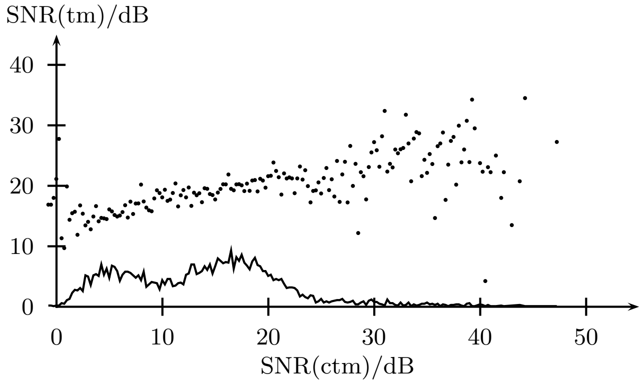

Figure 1: Comparison of SNRs for close-talk microphone (ctm) and throat microphone (tm) signals (*line: histogram of close-talk microphone SNR, dots: dependency of averaged throat microphone SNR on close-talk microphone SNR*).

the close-talk microphone signal. For less noisy signals (20dB and higher) the close-talk signal usually provides a clearer speech signal with a better frequency response resulting in higher SNR values compared to the throat microphone signal. A linear dependency between both SNRs can be approximated in Figure 1. The variance of the SNR values along the approximated line is rather high, which is noticeable for high and low SNRs where the variances were not averaged out due to a low number of measured values (see histogram in the same figure). Thus, a dependency between the signal-to-noise ratio of both signals exists but is quite weak.

Spectral Characteristics

Close-talk microphone signals sound more natural compared to signals from a throat microphone. This is due to limitations in the frequency range of the signal. While spectrograms of the lower frequency range are quite similar, the ones of the high frequency range differ substantially (compare spectrograms in Figure 2). Depending on the phoneme, frequencies from 4 kHz and above appear to be less present or entirely inaudible in the throat microphone signal. Thus, this limitation is due to the sensing kind and location of the transducer relative to the speech source. Apparently, the signal conducted by skin near the larynx (as in the case of a throat microphone) is of limited bandwidth.

Figure 3 shows two corresponding spectra of the consonants /*g*/ and /*s*/ for both microphones. The spectra have been computed using a manually selected window of the audible sound of each phoneme. In the very low and in the high frequency band both spectra for the phoneme /*g*/ appear to be very similar. But for the medium frequencies between about 1 and 5 kHz, the sound pressure level for the throat microphone is higher and the signal is more harmonic than the close-talk microphone signal. However, in the throat speech, the closure region of each of the voiced stops is characterised by distinct well-defined formant-like structures due to the placement of the throat microphone close to the vocal folds. This is accordant to Shahina and Yegnanarayana [6], who stated that voiced stop consonants like /*d*/ and /*g*/ are represented better in case of throat speech.

The spectra of the phoneme /*s*/ in Figure 3 shows, that

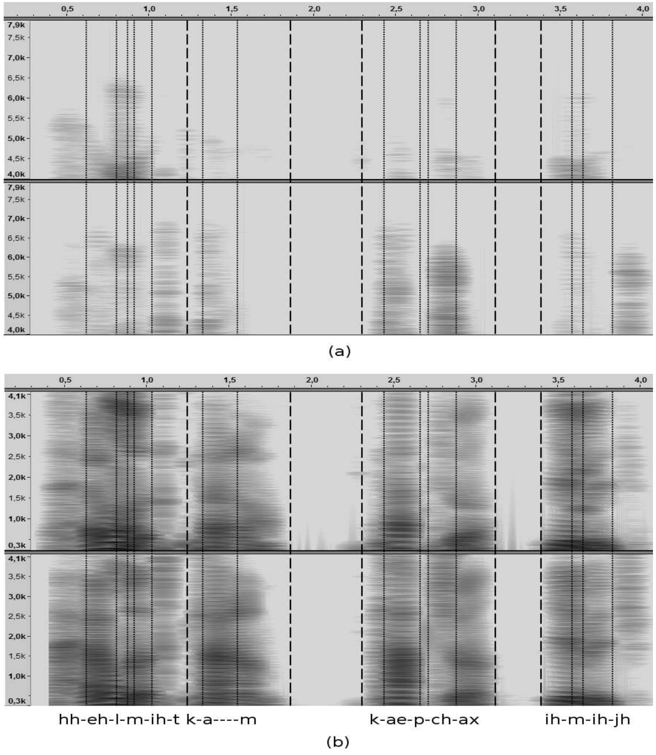

Figure 2: Comparison of spectrogram for 'clean' utterance *'HELMET CAM CAPTURE IMAGE'*. (a) high frequency band, (b) low frequency band. (*Top: throat microphone, bottom: close-talk microphone*)

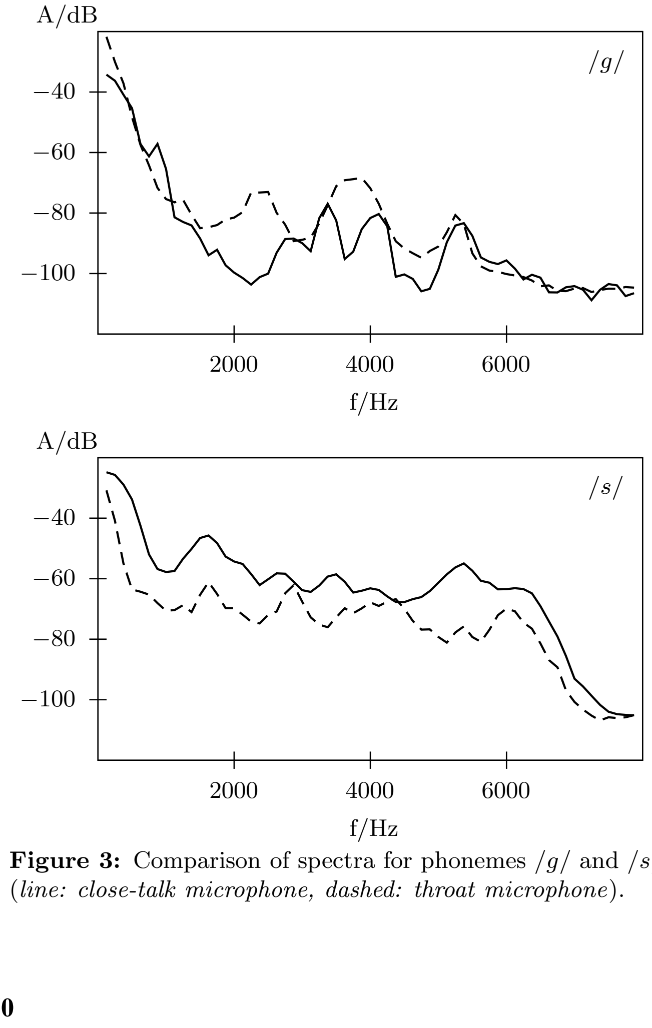

Figure 3: Comparison of spectra for phonemes /*g*/ and /*s*/ (*line: close-talk microphone, dashed: throat microphone*).

different phonemes are affected in different ways. While /g/ is better represented in the throat microphone signal, /s/ is stronger in the close-talk microphone signal. It is obvious that spectral differences between close-talk microphone and throat microphone are dependent on the phoneme as different parts of the vocal tract are responsible for the articulation of each phoneme. Analogously, Jou et al. [4] showed that a throat microphone signal cannot be modelled by simply applying a (sigmoidal) low pass filter because of this fact. Jou et al. claimed that consonants such as /m/ and /s/ differ significantly in both signals: The /m/ in a throat speech spectrum looks more like a vowel, while /s/, which is also strong at a high frequency and weak at a low frequency, is hard to hear in a throat microphone signal and subsequently also hard to recognise. Both assumptions could also be recognised in our evaluation.

Evaluation of ASR Performance

In the previous section, several differences between the signals of a close-talk and a throat microphone were stressed. Here, the effect of the different characteristics and concepts on the speech recognition performance is evaluated.

Evaluation Setup

The training of the acoustic models and the evaluation is performed by the Hidden Markov Model Tool Kit (HTK)[1]. Two sets of acoustic models are prepared for each microphone channel. First, clean speech models are trained based on about 1500 utterances recorded in a silent environment and, second, noisy speech models are trained on about the same number of utterances recorded on the motorcycle. The acoustic models are based on 12 mel frequency cepstral coefficients (MFCCs) with energy and first and second order derivatives. Cepstral normalisation is performed. Due to the very small amount of data, monophone models are trained. Speaker and environmental variability is taken into account by using 16 Gaussian mixtures.

Phoneme recognition without high-level syntactic relations is performed assuming an equal distribution of all phonemes to evaluate the acoustic effects. Two test sets are prepared, a clean speech test set with 417 clean utterances of two different speakers and a noisy speech test set of 588 noisy utterances recorded from three speakers while riding a motorcycle. Both test sets are neither a subset nor an intersection of the training set.

Evaluation Results

Figure 4 compares the phoneme accuracies for both microphone channels and different SNRs. Only the noisy test data is used grouped in 5dB steps by estimated SNR of the close-talk channel. Both sets of acoustic models are evaluated. The mismatch between clean training data and noisy test data is much less for the throat microphone signal than for the close-talk microphone signal as environmental noise has less impact on the throat microphone (which could also be seen by the higher SNR in Figure 1). Thus, the results of the recognition at low SNRs are much better for the throat microphone. For acoustic models based on noisy data (without any mismatch), the close-talk signal outperforms the throat signal due to the better frequency characteristics of the microphone.

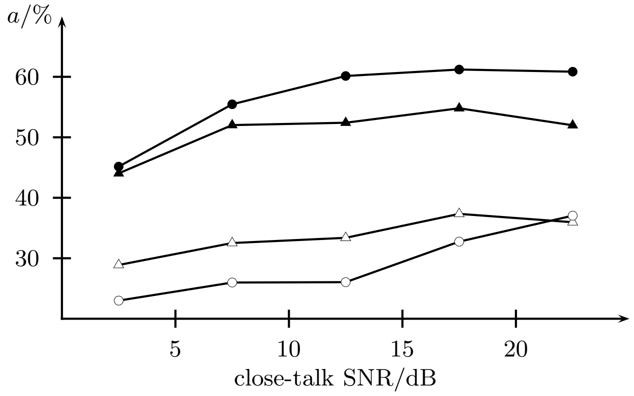

Figure 4: Phoneme accuracies $a/\%$ for close-talk microphone and throat microphone different acoustic models (△ - *throat microphone, o - close-talk microphone, filled symbols - noisy training data*)

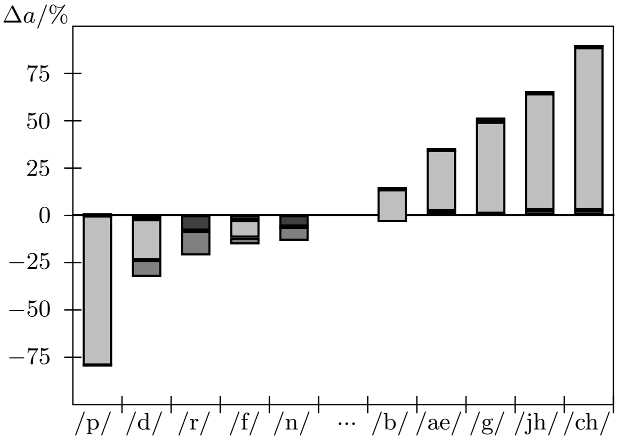

Figure 5: Differences of single phoneme accuracies $\Delta a/\%$ between throat microphone and close-talk microphone (*dark gray: phoneme correctness, light gray: insertions, gray: deletions*).

The bar chart in Figure 5 shows the absolute differences in single phoneme accuracies between throat microphone and close-talk microphone. The phonemes with the biggest absolute difference are presented. Negative values indicate worse accuracies for the throat microphone. The figures are estimated with clean speech models and clean speech test data only. Influences of phoneme correctness, insertion and deletion contributing to the overall phoneme accuracy are shown in different gray scales. Especially, /p/ is recognised with a significantly higher accuracy by the close-talk microphone mainly due to less insertions of this phoneme. A probable reason might be a similarity to non-speech sounds like gulps or other reflexes close to the larynx. The lack of typical characteristics of the plosive /p/ in the higher frequency bands makes the recognition of this phoneme also more

[1] http://htk.eng.cam.ac.uk/

difficult. The phonemes /*ch*/, /*jh*/ and /*g*/, on the other hand, are recognised better by the throat microphone. The difference is again mainly due to a difference in the number of insertions while phoneme correctness and number of deletions are almost equal. The phonemes /*ch*/ and /*jh*/ are similar to unharmonic noises like breathing and other non-speech airborne sounds, which might cause additional insertions when captured by the close-talk microphone. The phoneme /*g*/ is assumed to be generally easier to be recognised from the throat microphone signal as we already discussed in the previous section.

	ctm		*tm*	
	clean	*noisy*	*clean*	*noisy*
nasal	93.59	94.35	91.89	96.35
plosive	97.09	96.96	95.31	95.78
vowel	96.32	97.00	90.83	96.78
liquid	87.74	86.09	72.03	77.78
fricative	96.30	89.24	94.71	95.10
sil	92.64	91.81	97.17	98.21

Table 1: Specific phoneme group accuracy rates for close-talk (ctm) and throat microphone (tm) data. Clean test data is evaluated with clean speech models, noisy test data with noisy speech models.

A comparision of phoneme group accuracy rates is presented in Table 1. We roughly grouped all phonemes into five groups plus silence (for /*sil*/ and /*sp*/) depending on their particular articulation. Instead of the single phonemes certain phoneme groups containing the modeled phonemes should be recognised. Both test sets (clean and noisy) are evaluated using the appropriate clean or noisy speech models for each set. Liquids like /*l*/ or /*w*/ are generally recognised worse than other phonemes — especially using the throat microphone signal for recognition. For the close-talk microphone the accuracy rate of the fricatives drops significantly for noisy data. Fricatives usually suffer most from environmental noise due to their unharmonic, noise like characteristics. Thus, this problem does not occur for the throat microphone signal as it is less prone to environmental noise. For similar reasons silence can be recognised slightly better from throat microphone signals. Generally, phoneme group recognition on the throat microphone signal performs better for noisy speech than for clean speech. A reason could be the varying quality of clean speech data from the throat microphone. This data also includes several less intelligible utterances caused by a lack of pressure of the transducer to the larynx for some of the speakers.

Conclusion

The results of the evaluation show strength and weakness of throat microphones for automatic speech recognition. The robustness towards environmental noise and the different frequency characteristics of the throat microphone affect the ASR performance in several ways. The throat microphone signal is less influenced by environmental noise and, hence, performs well even for environmental mismatch between training data and test data. Without any mismatch the close-talk microphone usually performs better due to the better quality of the speech signal. While the throat microphone enables improved recognition results for some specific phonemes, recognition accuracy based on the throat microphone signal decreases significantly for others.

All in all, the results for both microphones are quite good for the difficult noise conditions on the motorcycle, though the high quality close-talk microphone performs slightly better. A clever combination of both microphone signals might further increase the overall performance. For even noisier environments the results for the ASR will most likely benefit from capturing speech by a noise robust throat microphone instead of the more vulnerable close-talk microphone. Here, further research will be necessary.

Acknowledgement

We would like to thank the European Commission for kindly funding the project MoveOn as well as all involved project partners for their work in the creation of the MoveOn database.

References

[1] The NIST SPeech Quality Assurance (SPQA) Package. Software - Version 2.3.

[2] ETSI ES 202 050, "Advanced front-end feature extraction algorithm", November 2003.

[3] Jingdong Chen, Jacob Benesty, Yiteng Arden Huang, and S. Doclo. New insights into the noise reduction wiener filter. *IEEE Transactions on Audio, Speech & Language Processing*, 14(4):1218–1234, 2006.

[4] Szu-Chen Jou, Tanja Schultz, and Alex Waibel. Adaptation for soft whisper recognition using a throat microphone. In *Proceedings of the International Conference on Speech and Language Processing, ICSLP 2004*, Jeju Island, Korea, October 2004.

[5] Bhiksha Raj and Richard M. Stern. Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine*, 22(5):101–116, September 2005.

[6] A. Shahina and B. Yegnanarayana. Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

[7] Thomas Winkler, Theodoros Kostoulas, Richard Adderley, Christian Bonkowski, Todor Ganchev, Joachim Köhler, and Nikos Fakotakis. The MoveOn Motorcycle Speech Corpus. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008.