An Investigation Into Incorporating Visual Information in Audio Processing

Ender Tekin, James M. Coughlan, and Helen J. Simon

The Smith-Kettlewell Eye Research Insitute, San Francisco, CA
{ender, coughlan, helen}@ski.org

Abstract. The number of persons with hearing and vision loss is on the rise as lifespans increase. Vision plays an important role in communication, especially in the presence of background noise or for persons with hearing loss. However, persons with vision loss cannot make use of this extra modality to overcome their hearing deficits. We propose automatically utilizing some visual information in hearing aids through the addition of a small wearable camera. Our initial results show potentially significant benefits to incorporating low level robust visual cues when the background noise is high. This technique can potentially benefit all persons with hearing loss, with substantial improvements possible for the speech perception performance of persons with dual sensory loss.

1 Introduction

In daily conversations, visual information significantly improves comprehension of speech [8], as well as providing the location and identity of the speaker. Seeing the speaker's face can account for up to 10 decibels improvement in speech perception in the presence of speech-babble noise [5]. Persons with hearing loss can use techniques such as speech-reading to significantly improve their speech perception. However, as the population in most developed countries is aging at a significant rate, the incidence of vision loss as well as hearing loss is also on the rise. Persons with such dual sensory loss cannot utilize the complementary information present in the facial cues, making communication more difficult.

The proportion of the population with vision and/or hearing loss directly correlates with age. Age related vision loss also manifests as loss in contrast/color sensitivity and temporal resolution, which cannot easily be corrected by lenses [4]. Age related hearing loss can also lead to decreased spectral and temporal resolution [4], making it harder to isolate the sound from a speaker in the presence of non-uniform noise, such as that due to other speakers in the environment, a problem known as the *cocktail party* effect. Such dual sensory loss affects more than 1 in 5 persons aged 70 years and over [6]. In this paper, we present our investigations into the benefits of utilizing the visual information in an intuitive and reliable way to enhance digital signal processing algorithms such as those used in current hearing aids. This research may lead to substantial benefits for persons with hearing loss, especially those with additional vision loss.

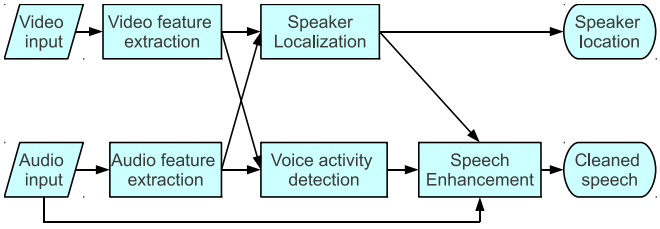

Fig. 1. Overview of the proposed system

Due to technological limitations, conventional hearing aids are unable to provide satisfactory performance in noisy and crowded environments [3]. Directional microphones can offer a 3-5dB improvement over omni-directional microphones when there is a known nearby speaker, such as at a restaurant [1]. Some microphones can even use a technique called beamforming to overcome the limitations of the static beam patterns of most hearing aids. However, there are often other sources of noise in the vicinity of the speaker that limit their effectiveness [2].

Most of these limitations may be alleviated, allowing significant performance improvements, by incorporating visual information in the audio processing stage. We propose integrating a wearable camera to obtain such visual information. This proposed system can improve the efficiency of speech processing algorithms, as well as provide information to the user regarding the identity of the speaker. There is similar, promising research into speech recognition algorithms that model head movements and lip articulations. However, these results are mostly limited to very well controlled conditions where the facial image is high resolution and well-lit. Such conditions are uncommon in the real world; as a result, we focus on using more robust cues, e.g., detection of mouth movements and the speaker in crowded environments using low-resolution videos.

2 Approach

Almost all speech enhancement techniques rely on being able to detect voice activity reliably; this allows these techniques to sample the background noise efficiently and frequently in order to enhance speech, as the statistics of noise are highly dynamic. In noisy environments, conventional algorithms can easily make mistakes since they only rely on the noisy audio cues to make their decisions. In these cases, using video of the speaker as well, voice activity may be detected more robustly; we illustrate this in Figure 1. In addition, since only certain acoustic components (e.g., signal power in different bands) will be related to a given speaker, it may be possible to learn these bands and inform the spectral speech enhancement algorithms to focus on eliminating sounds that are clearly unrelated to the speaker. We note that this proposed video information can supplement existing speech enhancement techniques, and we can build on the existing wealth of research on hearing aids and similar technologies developed

for persons with hearing loss. Location and speech activity of a speaker can be obtained more reliably from the visual information which is often less noisy than audio in real-world scenarios. We have performed some initial experiments to show the feasibility of this approach. We extracted audio and video features at 30 frames per second. The audio features include Mel-frequency cepstral coefficients, long-term spectral divergence (LTSD), and energy of the audio spectrum in the frame. Video features include variance and mean intensity, as well as spatio-temporal derivatives of the mouth region of the visible speaker. These are low-level cues that may be reliably extracted from lower-resolution video as opposed to the high-definition videos used in audio-visual speechreading research.

We determined the contributions of the various features by maximizing the mutual information of the linear combination of audio and video features, i.e.,

$$\boldsymbol{\alpha}^*, \boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} I(\boldsymbol{\alpha} \cdot \mathbf{f}_a; \boldsymbol{\beta} \cdot \mathbf{f}_v) \tag{1}$$

The results indicate a higher emphasis on video components such as the temporal statistics of the mouth area as captured by temporal gradients, and audio components such as LTSD and total audio signal energy in a window. We developed a simple audio-visual VAD using a linear combination of these features,

$$v = \boldsymbol{\alpha}^* \cdot \mathbf{f}_a + \boldsymbol{\beta}^* \cdot \mathbf{f}_v. \tag{2}$$

By comparing this value to a threshold τ, it is possible to determine whether voice activity is present. If $v > \tau$, voice activity is assumed to be present, and the speech enhancement algorithm attempts to clean the noisy speech signal. Otherwise, we assume that we are in a silence period, where the background noise estimate is updated by sampling the noise spectrum in this period.

3 Experimental Results

To establish the contribution of the visual information, a video-only VAD (where the audio components were nulled by setting $\boldsymbol{\alpha} = \mathbf{0}$) was compared with a conventional audio-only VAD based on the long-term spectral divergence [9]. The VAD outputs were used to enhance speech using a Wiener filter, [7], on a video with one speaker, where the audio signal was corrupted by various levels of additive white Gaussian noise and speech-babble noise. Segmental signal- to-noise ratio (SSNR) values were estimated to compare the performances as SSNR has been shown to correspond better with speech intelligibility than SNR.

Figure 2 illustrates this point. In this example, increasing levels of background noise were addded a quarter and a half-way through a video. While both approaches show similar performance for lower noise conditions, the performance of the audio-only method suffers more in high-noise situations, due to the inaccuracy of the voice activity detector at such high noise levels. In this case, the output SSNR using the video-VAD was on average 8dB better, showing that video can provide complementary information when audio is very noisy.

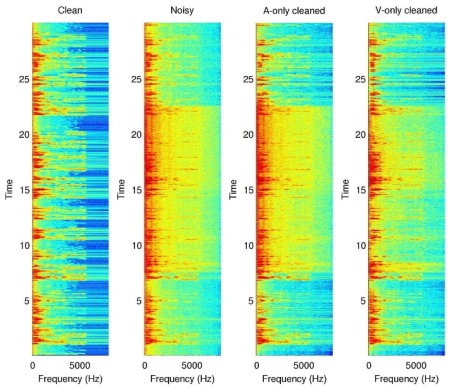

Fig. 2. Speech enhancement performance using video vs audio-based VAD's.

4 Conclusion

We are exploring the potential of incorporating video in speech-enhancement technologies. Our preliminary results are promising, and we plan to validate the output using subjective evaluations by persons with dual sensory loss. A system that includes video brings some challenges in usability, but the rising popularity and acceptance of wearable technologies, as well as the potentially significant improvements in speech perception performance may make this an attractive solution for the needs of a growing population with hearing and/or vision loss.

Acknowledgements. This research was supported by the NIH-NEI Award #R21EY022200, and the DoEd-NIDRR Grant #H133E110004.

References

1. Compton-Conley, C., Neuman, A., Killion, M., Levitt, H.: Performance of directional microphones for hearing aids: real world versus simulation. J. Am. Acad.Audiol. 15, 440–455 (2004)
2. Fabry, D.A.: Adaptive directional microphone technology and hearing aids: Theoretical and clinical implications. Hearing Review (Apr 2005)
3. Gatehouse, S.: Electronic aids to hearing. Br. Med. Bull. 63(1), 147–156 (2002)
4. Haegerstrom-Portnoy, G., Schneck, M.E., Brabyn, J.: Seeing into old age: vision function beyond acuity. Optom. Vis. Sci. 76, 141–158 (1999)
5. Helfer, K.S., Freyman, R.L.: The role of visual speech cues in reducing energetic and informational masking. J. Acous. Soc. Am. 117(2), 842–849 (2005)
6. Horowitz, A.: Dual sensory impairment among the elderly. Tech. rep., Lighthouse International and AARP Andrus Foundation (2001)
7. Loizou, P.C.: Speech Enhancement: Theory and Practice. CRC Press (Jun 2007)
8. MacLeod, A., Summerfield, Q.: Quantifying the contribution of vision to speech perception in noise. Br. J. Audiol. 21(2), 131–141 (May 1987)
9. Ramírez, J., Segura, J.C., Benítez, C., Torre, Á.D.L., Rubio, A.: A new adaptive long-term spectral estimation voice activity detector. In: Eurospeech. pp. 961–964. Korea (Oct 2004)