ROBUST HANDS-FREE VOICE CONTROL FOR MEDICAL APPLICATIONS

B.E. Sarroukh, L.C.A. van Stuivenberg, and C.P. Janse

Philips Research Laboratories
5656AE Eindhoven, The Netherlands
Email: {eddine.sarroukh, leon.van.stuivenberg, kees.janse}@philips.com

ABSTRACT

Hands-free voice control is an attractive solution for interaction in the medical environment. Considerable improvements in imaging technology have led to a continuously increasing demand for efficient methods for the manipulation of electronic images in medical applications. However classical interaction modalities, like remote control/mouse/keyboard, are either inadequate or impractical. To match the clean close-talk conditions where the speech engine is trained, microphone array processing is required. This paper proposes a robust speech enhancement scheme, for a hands-free voice control system for a medical X-ray application. This scheme combines microphone array processing and preprocessing in the speech engine into a single front-end, resulting in optimal accuracy. Artificial acoustical conditions as well as one realistic condition are investigated, to assess the feasibility of hands-free voice control for this application. The proposed solution enables the doctor to focus on the patient and images while interacting with medical equipment.

Index Terms— Voice control, speech recognition, beamforming, microphone array, noise suppression

1. INTRODUCTION

Voice recognition has become a widely accepted solution for controlling medical equipment. Often, it is the only feasible option in hands-busy situations or in environments where sterility is a priority. Other advantages of voice control are an increased freedom of user movement and the elimination of repetitive motions that would be needed for manual control.

Conventional voice control systems, however, have several shortcomings. In order to guarantee robust performance, it is necessary to place the microphone close to the mouth of the speaker. This requires the user to either wear a headset or use a clip-on microphone, which is often experienced as inconvenient and impractical, especially during procedures that may last for several hours. Microphone wiring or, alternatively, the implementation of a reliable wireless transmission system, can cause other practical problems when a close-talk microphone is used.

Deployment of a distant-talk microphone eliminates the need to wear a microphone and to establish a connection - wired or wireless - between the user and the device. However, in this case the microphone not only picks up the desired speech signal but also interfering signals such as background noise, sounds generated by the device itself, and speech signals that are not intended for controlling the device. In most cases, this will deteriorate the performance of the speech recognizer to an unacceptable level.

An array of microphones and adequate processing of the signals picked up by the microphones can improve the quality of the desired speech signal such that a speech recognizer can work properly. An appropriate combination of the microphone signals results in spatial filtering which suppresses all sounds coming from other directions than the desired speech signal. In this way, the speech signal controlling the device is enhanced and in many cases, the resulting signal is comparable to the output of a close-talk microphone.

After a short description of automatic speech recognition (ASR) in Section 2, the general challenges regarding robustness are described. In Section 3 we propose a solution scheme that we tested in practice. The evaluation of the implementation of this solution is given in Section 4. The solution proved to be robust against the acoustical changes in the examination room. Conclusions are drawn in Section 5.

2. AUTOMATIC SPEECH RECOGNITION

A system for automatic speech recognition typically consists of two stages: training and recognition. A model of speech production is the binding factor between these stages. A well-known choice of model for continuous speech recognition systems is based on the Hidden Markov Model (HMM), which is typically used for capturing the time-varying characteristics of phonemes. The collection of HMMs, corresponding to the different phonemes, describes the acoustical model. The parameters of the acoustical model are estimated during the training phase. In the recognition stage, the HMMs are concatenated according to the phonetic transcription provided in a lexicon of a word to construct a model of the complete word. The task of the recognition engine is then to estimate the most likely sequence of words that was spoken, given a segment of speech signal.

The robustness of a speech recognition system depends on several factors, a number of which are discussed below. The performance of command- and control-style applications is often indicated on two axes: the accuracy and the false-alarm rate. These are derived from the types of errors that can occur:

substitution *A command was spoken intentionally to the system, but the system recognized it as another command.*

deletion *A command was spoken intentionally to the system, but the system ignored it.*

insertion *No command was spoken to the system, but the system recognized a command anyway.*

The accuracy of a system is measured by the amount of substitutions and deletions, while the false-alarm rate is measured in number of insertions per hour of operational system listening time. Mismatch between the speech material used during training and the speech signal encountered during recognition is the major cause of substitution and deletion errors. Insertion errors on the other hand are mainly caused by so-called out-of-vocabulary speech, and non-stationary noise.

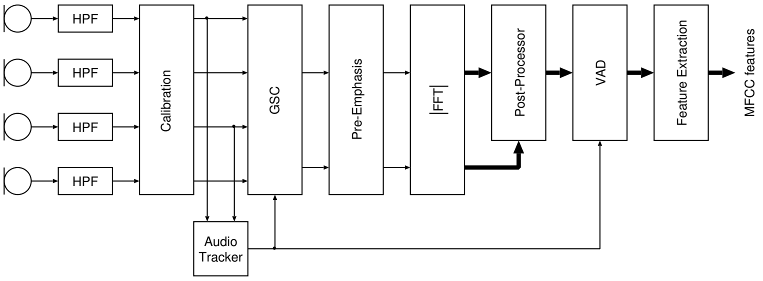

Fig. 1. Block diagram of the signal processing front-end: HPF: High-pass Filter; GSC: Generalized Sidelobe Canceller; FFT: Fast Fourier Transform; VAD: Voice Activity Detector; MFCC: Mel-Frequency Cepstrum Coefficients.

Difference between the acoustical conditions in which the training material was recorded and the conditions in which the recognizer is deployed can severely affect the recognition performance. The difference can often be attributed to the presence of noise and reverberation of the desired speech signal. The combined disturbances give rise to a nonlinear distortion in the speech feature domain, and consequently the features calculated from the noisy speech signal do not match the feature distributions contained in the acoustical model very well. In this paper we consider only speaker independent speech recognition and focus on the acoustic mismatch caused by the environmental acoustical conditions.

3. PROPOSED SOLUTION

In this section, we discuss the front-end of the hands-free system. The front-end is responsible for providing the speech recognition engine with a noise-free speech signal containing only the desired speech. The front-end does not feed a time-domain signal to the recognizer, but it computes features that ideally retain all information necessary for recognizing speech, while leaving out any unwanted information such as differences among speakers or background noise characteristics. Fig. 1 shows a block diagram of the signal processing front-end. The individual blocks in the processing chain are described in the following sections.

We use a linear array of length 30 cm, see Fig. 2, consisting of 4 microphones which we found to be a good compromise between performance and complexity. We make use of the total length of the array at low frequencies while minimizing spatial aliasing at high frequencies. All microphone signals are high-pass filtered before they are fed to the beamformer. This removes any DC offset and low-frequency noise. Calibration is used to compensate for gain differences between the individual microphone signals. These differences may stem from production tolerances, which can reach up to ±4 dB for cheap microphones, and from differences in the settings of the channels of the analog pre-amplifier.

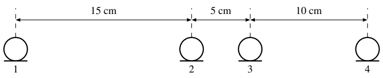

Fig. 2. Array configuration.

3.1. Audio Tracking

The audio tracker uses two of the four microphones to detect the angle at which the user is speaking. Different audio localization algorithms can be used, a detailed description of the algorithms can be found in [1], Chapter 4. The angle determined by the audio tracker can be used in different ways, depending on the implementation. It can be used by the beamformer to focus on the location of the desired speaker. In this case it must be guaranteed that the angle determined by the audio tracker correctly corresponds to the direction of desired speech. This could be done by speaker identification or by face recognition. In the medical scenario that we addressed, namely cardiovascular X-ray, the user is at a more or less fixed but unknown position because he/she is inserting a catheter into the patient. In this case, the position of the user can be determined from e.g. an activation command. The angle of any sound source detected by the audio tracker is then compared against this angle. The current angle of the sound source can also be used by the voice activity detector if the location of the desired speaker is known. In this case, any sound coming from a direction other than that of the desired speaker is identified as undesired and can be suppressed.

3.2. Generalized Sidelobe Canceller (GSC)

The structure of the beamformer is shown in Fig. 3. It is a generalized sidelobe canceller [2, 3] with a beamformer adapting to the sound source and an adaptive noise canceller W, implemented as a multi-channel frequency domain adaptive filter. The beamformer generates a primary output signal containing the desired signal z, and a set of noise references x that are fed to the noise canceller to cancel the remaining noise in the primary signal. The noise references x or, alternatively, the output signal of the noise canceller y are also used in the spectral post-processor discussed in Subsection 3.3. More information on beamforming and implementations can be found in [2, 3, 4, 5, 6, 7].

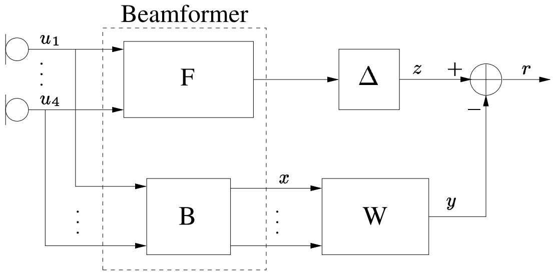

Fig. 3. Structure of the GSC.

3.3. Spectral Processing

A pre-emphasis filter (Fig 1) compensates for the natural attenuation of higher frequencies of voiced sections of speech. It is implemented as a first order finite impulse response filter:

$$r(n) = z(n) - \alpha z(n-1), \quad \alpha \in [0,1].$$

This filter is applied to the output of the beamformer as well as to the noise references x and to the output y of the noise canceller W for use in the spectral post-processor.

The time-domain signals, beamformer output, noise references and noise canceller output are transformed to the spectral domain by windowing overlapping blocks of input data with a Hamming window and computing an FFT. Only the magnitudes of the FFT coefficients are used in further processing. The spectral post-processor enhances the performance of the beamformer by increasing the suppression of interferences provided by the noise canceller W. It is based on spectral subtraction [8, 9] and it uses either the output of the noise canceller W or the noise references x from the blocking matrix B as an estimate of the interference. In the case that the noise references are used, the magnitudes of the FFTs of the different reference signals are summed in order to obtain a single interference estimate.

The postprocessing uses noise floor estimations based on minimum statistics. The noise estimation parameters, together with noise gain limitations and smoothness parameters need careful tuning for the specific recognizer to obtain optimal accuracy results. In this paper, we used default parameters and focused on the question whether the front-end leads to consistent improvement independent of the choice of the speech engine.

3.4. Voice Activity Detection (VAD)

Voice activity detection tries to separate time periods that contain potentially useful speech from all other periods. Only when voice activity is detected, should the input signal be passed on to the recognizer. Using a voice activity detector has two advantages: first, the computational complexity is decreased by recognizing speech only during periods of speech activity. Second, the risk of false recognitions is reduced by blocking out noise or undesired speech.

Our VAD is based on the detection used in the speech recognition system Phicos developed by the former Philips speech recognition research group in Aachen, Germany [10], which is based on two criteria: signal-to-noise ratio (SNR) and direction of arrival (DOA) of the sound source. The SNR is measured after the spectral post-processor and the DOA is provided by the audio tracker. The DOA can of course only be used as a criterion for the VAD if the location of the desired speaker is known.

Note that a VAD for speech recognition is usually not required to indicate speech activity per input frame, but it should assign speech activity to entire speech utterances which might consist of several words. Usually it is also required that the VAD opens some time before the utterance starts and only closes some time after the utterance has finished. This means that the VAD introduces some delay which may conflict with delay constraints of real-time systems.

In general, after applying the VAD, the signal is converted to the time-domain and fed into the speech engine. For Phicos, we disable the processing front-end of the engine and apply the feature extraction in the enhancement front-end directly after the VAD. Combining the preprocessing of the speech engine and the postprocessing leads in general to better recognition accuracy and avoids double enhancement.

3.5. Feature Extraction

In the feature extraction block, the frequency-domain data is transformed to features used by the speech recognizer. For the majority of state-of-the-art recognizers these features and possibly their derivatives over time. MFCCs are computed by reducing the frequency resolution of the FFT coefficients according to a mel scale[1] by applying a filter bank to the FFT coefficients. Each output channel of the filter bank is determined by computing a weighted sum of a number of FFT bins:

$$Y_m = \sum_{k=L_m}^{K_m+L_m-1} w_k^m X_k, \quad m = 0, \ldots, M-1, \qquad (1)$$

where Y_m is the output of the m^{th} filter bank channel, K_m is the number of frequency bins that are combined in this channel, L_m is the index of the lowest input frequency bin covered by this filter bank channel, X_k is the k^{th} FFT frequency bin of the input vector, w_k^m is the k^{th} element of the weighting function function of channel m, and M is the number of filter bank channels. We use triangular weighting functions. These triangular weighting functions are spaced evenly on a mel scale and are usually overlapping. The MFCCs are computed by taking the logarithm of the output signals of the filter bank and applying a discrete cosine transform (DCT) as shown in Fig. 4.

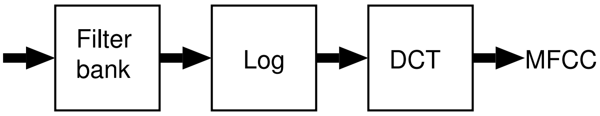

Fig. 4. MFCC Feature extraction from FFT coefficients.

4. EVALUATION

In this section we evaluate the effectiveness of the various signal enhancement steps of the scheme in Fig. 1. A comparison is made between applying no preprocessing but directly feeding the microphone signal to the engine (none), and the enhancements with the Beamformer (BF), the GSC, and the GSC combined with noise suppression (NS).

Various collections of speech material were gathered. In this paper we consider three of them. We recorded own data sets for the evaluation, named LAB, LAB-N and LAB-R. All these collections concern fairly simple command and control recognition tasks, using US English vocabularies. The command sets were recorded at various locations under varying noise conditions. The LAB data set represents office conditions where stationary noise from computer cooling fans is present. The reverberation time T_{60} was estimated at 400 ms. The microphone array was positioned approximately 1.5 m from the ground. The commands were spoken by eight non-native speakers of US English, whose positions/directions from the microphone array ranged from 1.2 m to 2.0 m / -45 deg to 45 deg.

The LAB-N data set is recorded in a hospital. The doors of the cabinets where the X-ray equipment was running were left open which increased the noise levels. Furthermore, background babble noise was present during recording. The recording scenario is similar to the LAB data.

The LAB-R data set represents a real-life scenario. The voice commands are given by the clinician, who is also occasionally talking to several other clinicians at the table side. Conversations with

[1] The mel scale attempts to map the perceived frequency of a tone onto a linear scale: $m = 2595 \log_{10}(1 + f/700)$, where m is the frequency in mel and f is the frequency in Hz.

Table 1. Performance on the clean LAB dataset.

	Phicos	C1	C2	MS6.1
none	92%	93%	88%	70%
BF	95%	95%	91%	80%
GSC	97%	95%	92%	83%
GSC+NS	96%	95%	93%	67%

the patient and between clinicians occur as well, with a few occurrences where the clinicians talk to people in the control room. The reverberation time T_{60} was estimated to be 550 ms. Due to the short distance (approximately 1 m) between clinician and microphones, the recognition is not expected to be significantly affected by speech reverberation.

We considered four speech engines: Phicos, C1, C2, and MS6.1. Phicos [10] is a continuous speech recognizer developed at Philips Research Aachen. It is a research tool used in the development of new algorithms for speech recognition. C1 and C2 are commercially available speech recognition engines, both supporting command and control applications. Various versions of the Microsoft Speech Recognizer exist, and are bundled with e.g. Microsoft's Speech SDK 5.1 (version 5.1), Office 2003 (version 6.1), and Windows Vista (version 8). We used version 6.1.

We note that no specific parameter tuning is applied to the different combinations of the algorithms with speech engines. The recognition accuracy on data set LAB is given in Table 1. Linear spatial filtering with the beamformer and GSC provides consistent improvements across all engines in terms of accuracy. Improving the SNR can lead to an increase in insertion errors, as we have observed during experiments. The spectral postprocessing does not prove to be very useful for the low noise LAB dataset, which can be attributed to the introduced signal distortion.

Table 2. Performance on the LAB-N dataset with various enhancement techniques.

	Phicos	C1	C2	MS6.1
none	19%	54%	74%	25%
BF	48%	75%	87%	40%
GSC	72%	81%	86%	56%
GSC+NS	81%	84%	89%	67%

The improvements obtained with signal enhancement if a reasonable amount of stationary noise is present (LAB-N dataset) are given in Table 2. Very significant improvements can be observed across all engines. The largest improvement is found for Phicos, which could be explained by the presence of single channel noise reduction methods in the commercial engines. The results of the various processing schemes are somewhat irregular for the C2 engine, which suggests that the sub-optimality of the simple concatenation of preprocessing and engine front-end are most notable here. Again, the beamformer and GSC provide consistent recognition improvements, and in most cases the spectral methods boost the performance very significantly.

For the LAB-R the results obtained with the various processing schemes are listed in Table 3. The most consistent improvements are obtained for the Phicos engine. The BF and GSC furthermore increase accuracy for the C1 and C2 engines; marginal effects are observable for the MS6.1 engine. The spectral methods generally give improved results, except for the C1 engine.

Table 3. Performance on the LAB-R dataset.

	Phicos	C1	C2	MS6.1
none	77%	78%	57%	85%
BF	85%	85%	63%	85%
GSC	87%	86%	63%	85%
GSC+NS	89%	86%	66%	87%

5. CONCLUSIONS

Substantial improvements in recognition performance can be attained with the proposed solution based on acoustical beamforming. Spatial filtering with beamforming provides consistent gains in recognition accuracy. Methods based on spectral subtraction in general suffer from distortion of the desired speech signal and thus are less suitable in relatively noise-free conditions. They do however provide very impressive results in the more noisy conditions, although care must be taken when integrating such a method directly with a complete speech recognition system, i.e. performing noise reduction twice should be avoided.

6. REFERENCES

[1] E. Hänsler and G.U. Schmidt, *Topics in Acoustic Echo and Noise Control*, Springer Verlag, 2006.

[2] L.J. Griffiths and C.W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beamforming," *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[3] W. Herbordt and W. Kellermann, "Computationally efficient frequency-domain robust generalized sidelobe canceller," in *Proc. Intl. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2001.

[4] S. Affes and Y. Grenier, "A Signal Subspace Tracking Algorithm for Microphone Array Processing of Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 5, pp. 425–437, Sept. 1997.

[5] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. on Signal Processing*, vol. 47, no. 10, pp. 2677–2684, 1999.

[6] W. Herbordt and W. Kellermann, "Frequency-domain integration of acoustic echo cancellation and a generalized sidelobe canceller with improved robustness," *European Trans. on Telecommunications*, vol. 13, no. 2, pp. 123–132, 2002.

[7] J. Benesty, C. Jingdong, H. Yiteng, and J. Dmochowski, "On microphone-array beamforming from a mimo acoustic signal processing perspective," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1053–1065, 2007.

[8] S.F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, pp. 113–120, Apr. 1979.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean–square error short–time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[10] V. Steinbiss et al., "The Philips Research System for Continuous-Speech Recognition," *Philips Journal of Research*, vol. 49, no. 4, pp. 317–352, Dec. 1995.