Use of a CSP-based voice activity detector for distant-talking ASR

Luca Armani, Marco Matassoni, Maurizio Omologo, Piergiorgio Svaizer

ITCirst - Centro per la Ricerca Scientifica e Tecnologica
Via Sommarive, 18 - 38050 Povo(Trento), Italy

larmani@itc.it, matasso@itc.it, omologo@itc.it, svaizer@itc.it

Abstract

This paper addresses the problem of voice activity detection for distant-talking speech recognition in noisy and reverberant environment. The proposed algorithm is based on the same Cross-power Spectrum Phase analysis that is used for talker location and tracking purposes. A normalized feature is derived, which is shown to be more effective than an energy-based one. The algorithm exploits that feature by dynamically updating the threshold as a non-linear average value computed during the preceding pause. Given a real multichannel database, recorded with the speaker at 2.5 meter distance from the microphones, experiments show that the proposed algorithm provides a relevant relative error rate reduction.

1. Introduction

Robust multi-sensory based interfaces represent a key issue for the development of perceptual systems able to understand a given audio-visual scenario. Future development of the related technologies is envisaged for applications in contexts as meetings, video-conferencing, interviews, etc.

This study is being conducted under the project PEACH (Personal Experience Active Cultural Heritage)[1], whose objective is to develop and experiment advanced technologies that can enhance cultural heritage fruition in a museum. Among many other objectives, the project refers to the realization of multisensory analysis for tracking a single visitor and recognizing distant-talking speech utterances.

Distant-talking ASR in a noisy and reverberant environment represents a difficult problem. In a realistic situation, one has to take into account several factors, generally neglected in a close-talking interaction scenario as, for instance, the fact that talker's position and his/her head orientation may be unknown and time-varying in an unpredictable fashion. The trend is that of approaching the "disappearing computing" concept, but this may also cause a speaking style sometimes difficult to characterize, rich of spontaneous speech phenomena as well as of rapidly variable speech dynamics.

Given these observations, it is necessary that the ASR front-end (either based on a single far microphone or on a microphone array) incorporates very robust processing techniques, even based on auditory perceptual cues. A preliminary step regards the possible introduction of an activity detection module able to distinguish among speech and other (stationary or not) noise sequences. Nowadays, in many applications of the ASR technology voice activity detection is accomplished inside the recognition process, that is without a two-step procedure. This requires a very good statistical acoustic modeling of noise and distortion preceding and following the given utterance. The alternative two-step procedure relies on the accuracy of the voice activity detector; some effective solutions are described also in the most recent literature [1, 4, 6, 10]. However, so far a few works addressed its impact on ASR performance in a distant-talking scenarios. In [9] adaptive energy thresholds were applied to the output of the delay and sum beamformer to identify speech boundaries. In [7], a frequency-domain processing led to the introduction of a coherence measure between two input channels that was used to detect a generic acoustic event. The application of a similar feature to speech/noise classification is also documented in [2].

In general, in distant-talking scenarios a reduced processing effort and an ASR performance improvement are envisaged when a reliable speech activity detector will be available. The scenario being investigated under PEACH comprises the distribution in space of several microphone pairs, providing signals at sample level synchronous to each other, which allows to locate and track the speaker position. The latter goal is accomplished by applying a very accurate Cross-power Spectrum Phase (CSP) based time delay estimation technique, that was developed in the past for surveillance and video-conferencing purposes [8]. In [7] it was evidenced that the CSP analysis might be useful also for speech detection activity purposes. Hence, a very appealing solution is envisaged that consists in the use of the same technique for both talker location and speech activity detection.

Another relevant aspect regards the impact of the resulting technique when a recognizer of medium-high perplexity, not trained in the given experimental context, is being used. Our previous work [5] addressed the problem of training a distant-talking ASR based on multi-microphone input, when a specific database is not available. In general, the recognizer may fail either when a too long non-stationary noise sequence precedes and follows the given utterance, or when the speech sequence is truncated at its "ideal" boundaries.

This preliminary study aims at providing a first comparison between a traditional energy-based technique and the here proposed one, for isolated word recognition task of medium vocabulary size (i.e. 200). Experiments were conducted in noisy and reverberant conditions which are typical of a large office or of a museum room.

The paper is organized as follows: in Section 2 the Cross-power Spectrum Phase analysis (CSP) is presented. In Section

[1]This work was partially funded by the Province of Trento, Fondo Unico PEACH project

3 the VAD based on CSP Coherence Measure is derived and a short presentation of the recognition engine here adopted is given. Section 4 includes experimental results and a related brief discussion, and finally Section 5 introduces to some conclusions and to the future work.

2. CROSS-POWER SPECTRUM PHASE ANALYSIS

The discriminating feature here discussed is a coherence measure between the signals of two microphones, that is the phase correlation computed by a Cross-power Spectrum Phase analysis. Using at least two microphones the capability of discerning between directive sources and spatially diffuse disturbances can be exploited. Moreover, the wide-band content typically characterizing speech is emphasized in contrast to narrow-band noise sources, that are deemphasized thanks to the properties of the given technique.

The procedure [8] for estimating a CSP-based Coherence Measure (CSP-CM) starts from the computation of spectra $S_1(t,f)$ and $S_2(t,f)$ through Fourier transforms applied to windowed segments of signals s_1 and s_2, centered around time instant t. Then, these spectra are used to estimate the normalized Cross-power Spectrum:

$$\phi(t,f) = \frac{S_1(t,f)S_2^*(t,f)}{|S_1(t,f)||S_2(t,f)|} \quad (1)$$

that preserves only information about phase differences between s_1 and s_2. Finally, the inverse Fourier transform of $\phi(t,f)$ is computed:

$$C(t,\tau) = \int_{-\infty}^{+\infty} \phi(t,f)e^{j2\pi f\tau}\,df. \quad (2)$$

The resulting function (of the lag τ) is the transform of an all-pass function and has a constant energy, mainly concentrated on the inter-channel delay δ.

The graphical representation of the CSP-CM allows one to make evident the mutual delay during the acoustic event emission, an interesting aspect being the intrinsic normalization of the resulting feature.

This behaviour gives the opportunity to exploit the CSP technique in an effective way also for VAD purposes.

Figure 1 illustrates an example of this analysis. The upper plot depicts the noisy speech signal acquired by a single microphone of a linear array. The lower plot represents the corresponding phase correlation between two channels of the array as a function of time (horizontal axis) and mutual delay in samples between the channels (vertical axis). A darker gray level denotes higher coherence.

Each sequence includes pauses (between words or in correspondence with stops) during which a direct waveform is not present. The corresponding small interruptions in each line show the accuracy and rapid adaptation of the CSP technique to abrupt changes of signal characteristics.

3. DISTANT-TALKING SPEECH RECOGNITION

A sizeable body of work on distant-talking automatic speech recognition have been produced in recent years [7]. The use

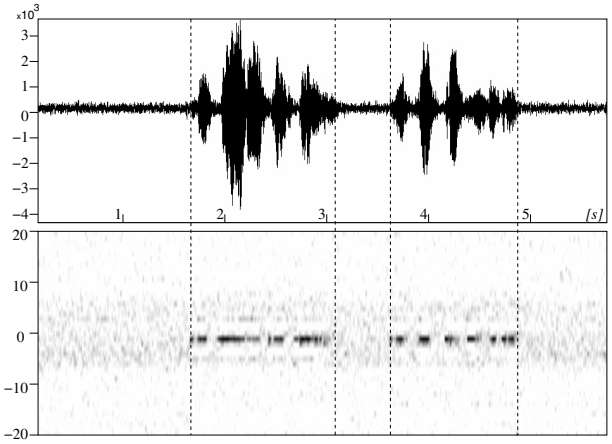

Figure 1: Example of CSP analysis applied to a microphone pair; the lower plot represents the phase correlation between the two channels. Vertical axis indicates the inter-channel delay. A darker gray level denotes higher coherence.

of either single microphones or multi-microphone systems has focused primarily on experimental contexts (e.g. car environment, teleconferencing room) for tasks generally characterized by a small-size vocabulary and by a low complexity language.

In this work the availability of a multi-channel input is exploited to study the impact of microphone distance on the voice activity detection and on ASR performance. In other words, no beamforming or other multichannel processing is here investigated. The recognizer front-end operates on a single microphone signal and the speech degradation associated to the distant-talking modality is compensated directly by a suitable training of the acoustic models, as explained in Subsection 3.2.

3.1. CSP-based Voice Activity Detection

Given the CSP-CM calculated on a microphone pair, a scalar value representing speech activity is obtained on the basis of the non-linear processing described in the following.

The maximum CSP-CM value of the current frame is compared to the current threshold value to detect intervals in which coherent directional wavefronts are arriving at the two microphones. The threshold is dynamically updated by calculating it as a nonlinear average value of CSP-CM amplitude during speech absence. More in detail, the most recent CSP-CM values of non-speech intervals are buffered and resorted in ascending order. The average value of the lower fraction (e.g. the lower half) of the reordered buffer is taken as the new current threshold. Potential speech segments are determined when the threshold is exceeded. Speech intervals are detected only when a set of conditions are satisfied by CSP-CM amplitude with respect to the dynamic threshold, i.e.

- the detected candidate segment is long enough ($> minlength$)
- inside the candidate segment CSP-CM values are below threshold only for short intervals ($< maxgap$)
- a sufficient percentage ($> mindensity$) of frames inside the candidate segment is over the threshold.

Here $minlength$, $maxgap$ and $mindensity$ are parameters characterizing the detection procedure. Figure 2 reports an example of the behaviour of the CSP-CM based VAD.

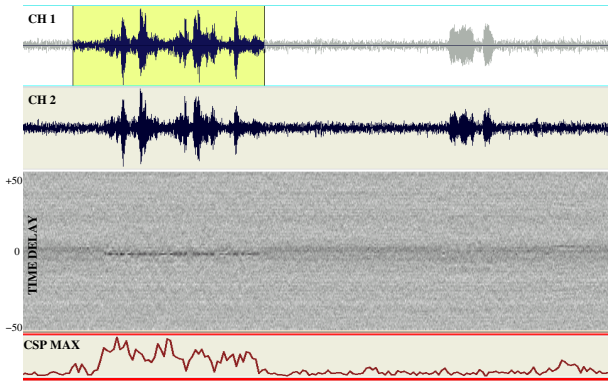

Figure 2: Example of CSP-CM based VAD: in the first channel of the pair is highlighted the resulting speech segment. The last plot shows the peak of the CPS-CM over time. The possible second candidate segment is a phone ring, correctly "rejected" by the VAD.

3.2. Recognition engine

The front-end processing here consists in a traditional MFCC feature extraction. The input signal is preemphasized and blocked into frames of 20 ms duration (with 50% frame overlapping). For each frame, 12 Mel-frequency Cepstral Coefficients (MFCCs) and the log-energy are extracted. MFCCs are normalized by subtracting the MFCC means computed on the whole utterance. The log-energy is also normalized with respect to the maximum value in the utterance. The resulting MFCCs and the normalized log-energy, together with their first and second order time derivatives, are arranged into a single observation vector of 39 components.

Acoustic modeling is based on a set of 34 phone-like speech units. Each speech unit is modeled with left-to-right continuous density HMM with output probability distributions represented by means of mixtures having 16 Gaussian components with diagonal covariance matrices. Context independent phone HMMs are trained either using a clean speech database or a noisy version, obtained by *contamination*.

A *contaminated* database consisting of acoustically realistic signals has been artificially recreated using a clean corpus along with knowledge (e.g. room impulse responses and background noise signals) of the real operating environment (see Figure 3). For this purpose, a simplified additive/convolutive model has been adopted as follows:

$$s_{co}(t) = h_r(t) \star s_{cl}(t) + k \cdot n(t) \qquad (3)$$

where $h_r(t)$ is an impulse response of the room, k is a scaling factor, $n(t)$ is background noise acquired in the room, s_{cl} is the clean speech, s_{co} is the contaminated speech, and $\star$ denotes convolution. The effect of background noise is accounted for by scaling the noise recorded inside the room using an appropriate amplitude to reproduce different SNR's and then adding the result to reverberant speech. The reverberation effect of the room was achieved by convolving the close-talking signal with impulse responses measured using a time-stretched pulse, as discussed in [5].

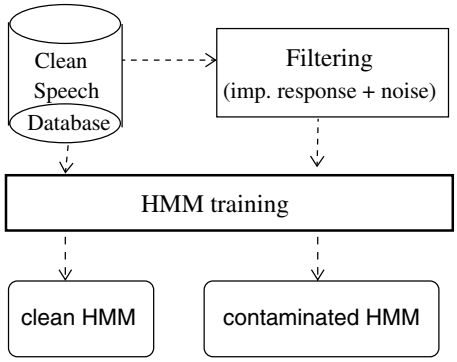

Figure 3: Data contamination procedure: a clean corpus is filtered taking into account the background noise and the impulse responses associated to several talker-microphone pairs in the room under test.

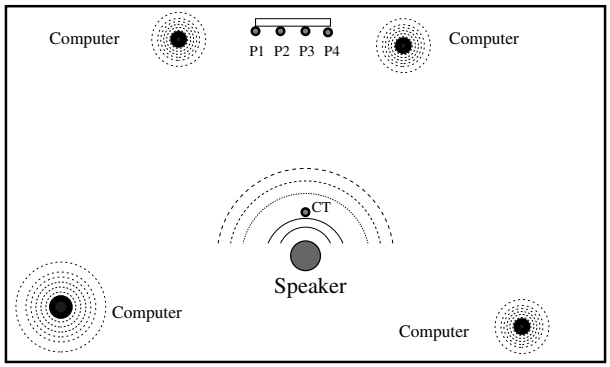

Figure 4: Plot of the room where the multi-channel corpus has been collected. The linear array is formed by four omnidirectional microphones P1, P2, P3, P4 and the speaker utters the isolated words in front of the array at about 2.5m. The noise is mainly due to computer fans and air conditioning.

4. EXPERIMENTAL RESULTS

To provide some performance figures a specific noisy corpus has been acquired. The multi-channel corpus consists of speech material collected in an office of size 7.0 m × 3.5 m × 3.5 m characterized by a moderate amount of reverberation ($T_{60} \simeq 0.3s$) as well as by the presence of coherent noise due to secondary sources (e.g. computers, air conditioning, etc.). Recording of each utterance was accomplished by using both a close-talking (CT) directional microphone and a linear array as shown in Figure 4. The four microphones of the array, spaced by 15 cm, are indicated by the labels P1, P2, P3, P4.

Standing in frontal position at about 2.5 m from the array, each speaker has uttered a sequence of isolated commands in-

serting among them a silence of variable duration from one to three seconds.

The main statistics of the database under test are reported in Table 1.

# spk.	dict. size	# words	rec. time	SNR
10	200	375	22 min	7 dB

Table 1: Statistics of the real multi-channel database: number of speakers, dictionary size, total number of words, recording time, and mean SNR of the sessions.

VAD	ideal	energy	CSP-CM
CT	4.5	5.6	-
P1 [15cm]	24.8	46.3	33.1
P2 [15cm]	24.8	50.7	33.6
P3 [15cm]	22.7	42.9	31.2
P3 [30cm]	22.7	42.9	32.5
P4 [15cm]	24.3	40.5	32.5
P4 [45cm]	24.3	40.5	33.1

Table 2: Recognition results in terms of WER (%). The column *ideal* represents performance obtained exploiting manual segmentation. In square brackets the distance between the microphones of the pair is indicated.

The performance given in Table 2 in terms of Word Error Rate (WER) confirms the advantage of this technique with respect to a classic energy-based one [3] that is used in all the on-field applications of the ITCirst's ASR technology. Despite the simplicity of the recognition task (long sequences of isolated words), WER on the *CT* microphone is 4.5% due to the presence in the vocabulary of many words very similar from the acoustic-phonetic point of view (e.g. dello/della, il/in, presidente/presidenza), which makes the system more sensitive to the VAD performance. It is worth noting that using a manual segmentation the error rate increases to about 25%. At the moment, that result represents the reference ideal performance. Given that upperbound, a detailed comparison between the two VAD systems show that using the CSP-CM based one a relative reduction rate of about 50% was obtained. Nevertheless, the performance loss between *ideal* and CSP-CM based VAD suggests to integrate the coherence function with other acoustic features. This topic is under investigation.

5. CONCLUSIONS AND FUTURE WORK

In distant-talking speech recognition the design of a reliable VAD module is crucial: the introduction of a new feature derived from a Cross-power Spectrum Phase analysis may successfully be adopted in order to improve recognition performance. The comparison with a standard energy-based VAD on a simple recognition task has shown the advantages of the proposed algorithm.

Further studies will address different issues among which the possibility of exploiting some power spectrum information that is neglected using the present CSP-CM analysis, the automatic tuning of the parameters of the detection algorithm, and the maximum distance of a microphone pair. The latter aspect is very important in order both to reduce the number of microphones in the given environment and to ensure a good monitoring coverage of the physical space.

Finally, given a large multichannel database (at this moment under collection) more effective VAD parameter methods will be explored in order to make the system self-adapting and in order to explore probabilistic schemes aimed at learning the parameters (e.g. threshold adaptation speed, analysis window) in a statistical fashion.

6. References

[1] S. Bou-Ghazale, K. Assaleh, "Robust Endpoint Detection of Speech for Noisy Environments with Application to Automatic Speech Recognition", *Proc of ICASSP*, 2002

[2] R. L. Bouquin-Jeannes, G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system", Speech Communication, vol. 16, pp. 245-254, 1995.

[3] G. Carli, R. Gretter, "A Start-End Point Detection Algorithm for a Real-Time Acoustic Front-End based on DSP32C VME Board", Proceedings of ICSPAT 92, Boston, USA, 1992.

[4] R. Hariharan, J. Hakkinen, K. Laurila, "Robust end-of-utterance detection for real-time speech recognition applications", *Proc. of ICASSP*, 2001.

[5] M. Matassoni, M. Omologo, D. Giuliani, P. Svaizer, "HMM-Training with Contaminated Speech Material for Distant-talking Speech Recognition", Computer Speech and Language, 16, pp. 205-223, 2002.

[6] E. Nemer, R. Goubran, S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain", IEEE Trans. Speech Audio Processing, vol. 9, pp. 217-231, Mar. 2001.

[7] M. Omologo, M. Matassoni, P. Svaizer, "Speech Recognition with Microphone Arrays" in Microphone Arrays, edited by M. Brandstein and D. Ward, Springer, 2001.

[8] M. Omologo, P. Svaizer, "Use of the cross-power-spectrum phase in acoustic event location", IEEE Trans. on Speech and Audio Processing, 5 3, pp. 288–292, 1997.

[9] D. Van Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings", Proc. of ICASSP, pp. 833–836, 1990.

[10] Q. Zou, X. Zou, M. Zhang, Z. Lin, "A robust speech detection algorithm in a microphone array teleconferencing system", *Proc of ICASSP*, pp. 3025–3028, 2001.