Speech Recognition in reverberant environments using remote microphones

Luca Brayda, Christian Wellekens
Institut Eurecom,
2229 Route des Cretes,
06904 Sophia Antipolis, France
{brayda,wellekens}@eurecom.fr

Marco Matassoni, Maurizio Omologo
ITC-irst,
via Sommarive 18,
38050 Povo (TN), Italy
{matasso,omologo}@itc.it

Abstract

This paper addresses distant-talking speech recognition by means of remote sensors in a reverberant room. Recognition performances is investigated for different ways of initializing, steering, and optimizing the related beamformer. Results show how much critical that front-end processing may be in such a challenging setup, according to the different positions and orientations of the speaker.

1 Introduction

Distant-talking speech recognition is a very challenging topic. To tackle it, microphone arrays [1] are generally employed thanks to the capabilities of beamforming techniques to enhance the speech message, while attenuating undesired contributions of environmental noise and reverberation. Microphone arrays can be steered toward the most convenient *look* direction, which ensures the best speech recognition performances. This can be accomplished by adopting a suitable filter-and-sum beamforming [2, 3], i.e. a combination of filtered versions of all the microphone signals. In the past, a wide literature addressed beamforming mainly with the target of deriving an enhanced signal with good properties from the perceptual point of view rather than maximizing speech recognition performances. More recent works have addressed the task of improving recognizer accuracy, which can represent a quite different objective. To this regard, a technique that deserves to be mentioned is Limabeam [4], which aims to optimize the beamformer parameters, given the most likely HMM state sequence that has been observed in a first processing step.

Moreover, an intensive activity of evaluating performances of microphone array based speech recognizers is being conducted world-wide, in particular in the communities related to the EC AMI and CHIL projects: NIST has recently organized benchmarking campaigns (see http://www.nist/gov/speech) which showed that the error rate provided by a 64-microphone array based recognizer is about twice the error obtained on the corresponding close-talking microphone signal, given a large vocabulary spontaneous speech recognition task. We observe that, when dealing with a real reverberant environment, the direction that ensures the best automatic speech recognition (ASR) performances can be different from the one determined by speaker localization techniques. In the past, accurate time delay estimation methods and related speaker localization systems were addressed which can be used to select a possible steering direction. However, also given this approach in a real-world situation one may encounter problems due to the head orientation that represents another source of variability very difficult to address: in other words, when the speaker is not aiming toward the array, the speech captured by each microphone of the array will be mostly characterized by contributions due to reflections. This paper investigates on distant-talking speech recognition in a real highly reverberant environment given different speaker positions, in most of the cases not oriented toward the microphone array. Existing techniques are presented and some new possible improvements are proposed. The purpose of the work is: to describe the parameters of a general microphone array processing system (Section 2), focusing to the beamforming techniques; to outline the possible performances that can be obtained steering the array in different directions (Section 3); to understand the potential of delay-and-sum beamforming, given delays extracted by a technique typically used for speaker localization purposes (Section 4); to outline the room for improvements estimating "recognition-oriented" filters (Sections 5 and 6) or exploiting additional information related to the environment such as the room impulse responses (Section 7). Finally, Section 8 describes the experimental setup and results (derived by using a multi-microphonic version of the well known TI connected digit recognition task) and Section 9 draws our conclusions and discussions for future work.

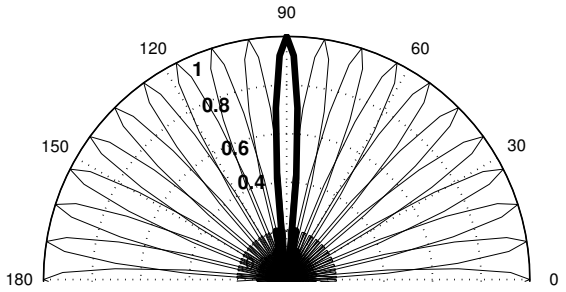

Figure 1. *Amount of the pi-space spanned by a microphone array with M=8, d=0.04m, f_{max}=7500 Hz and steered for 19 different angles. The main lobe of a single beampattern appears in bold, while the sidelobes, not plotted, are negligible.*

2 Microphone Arrays for ASR

Microphone arrays can be effectively used to improve the quality of speech signals by steering the array toward a specific look direction. Because a linear microphone array is a sampled version of a theoretical continuous sensor, the superposed response which approximates the corresponding continuous aperture response is a function of both the frequency of the received signal and its direction. The function, called *directivity pattern*, can be represented as:

$$D(f,\theta) = \sum_{m=0}^{M-1} W_m(f) e^{j\frac{2\pi f}{c} md\cos\theta} \quad (1)$$

where f denotes frequency, θ is the angle of arrival of signals in radiants, relative to the array axis, M is the number of microphones, $W_m(f)$ is the complex weight for sensor m, c is the sound speed and d is the inter-microphone distance. The main lobe of the directivity pattern is as much narrow as the frequency or d increase. If d exceeds half the signal minimum wavelength, spatial aliasing occurs. Expressing the array output as the as the sum of weighted channels, we have:

$$X(f,\theta) = \sum_{m=1}^{M} W_m(f) S_m(f) e^{j\frac{2\pi f}{c} md\cos\theta} \quad (2)$$

where $S_m(f)$ is the frequency domain signal received at at the m-th microphone and $X(f,\theta))$ is the output of the beamformer. Note that the output is equal to the directivity pattern if the received signals are equal to 1. In this work we focus on finding the set of parameters that shape the directivity pattern so that the recognition features extracted from $X(f,\theta)$ give the highest recognition rate, and not just the SNR, as possible.

3 Delay-and-sum and angle-driven beamforming

The simplest way to beamform multi-channel signals is Delay and Sum beamforming [5], i.e. when the weights $W_m(f)$ in Equation (1) are equal to 1. The aim is to set the delays $\tau_m = \frac{md\cos\theta}{c}$ for each microphone. Being the purpose to form a beam at a specific direction, given θ, a different set of delays can be calculated for each desired angle. In this work we focus on spanning the pi-space and look at equi-angled directions. For each direction we "steer" the array to a specific angle θ, then beamforming (theta-D&S) and recognition are performed: this results in getting a Recognition Directivity Pattern (RDP), the main lobes of which will "point" to regions where signals are better recognized. In order to cover all the space in front of the array, while avoiding aliasing, we propose to limit the number of beams R to:

$$R = \frac{\pi}{\arg_\theta D_{-3dB,l}(f_{max},\theta) - \arg_\theta D_{-3dB,r}(f_{max},\theta)} \quad (3)$$

where f_{max} is the maximum frequency of interest and the denominator is the main lobe width when the lobe attenuation is $-3dB$, which is the distance in radiants between the point to the left $D_{-3dB,l}$ and to the right $D_{-3dB,r}$ of the main lobe peak at -3dB. Thus, steering the array results in beamforming as depicted in Figure 1, where we considered 8 microphones, with 4 cm inter-microphone distance and a maximum frequency of 7500 Hz. This setting ensures aliasing to be negligible in the speech band. D&S generally performs better in environments where speech is affected by additive noise rather than reverberation, because it exploits the destructive interference of noise sources, which are generally uncorrelated to the source of interest. However, in reverberant environments the main noise source is the speaker himself. In our experiments we study the impact that reflections have on Word Recognition Rates (WRR) by observing the angles at which the RDP is higher.

4 Beamforming via Time Delay Estimation

The delays τ_m can also be estimated automatically. In very reverberant environments it is not trivial to estimate the inter-channel delays and perform D&S, because reflections behave like multiple highly correlated speech sources. The easiest approach to perform TDE between two microphones is the maximization of the value assumed by the cross-correlation as a function of the time lag. The correlation can be calculated as inverse Fourier transform of the cross-power spectrum $G_m(f) = S_m(f)S_r(f)^*$, (a given microphone r can be the reference for any pair, e.g the central microphone). In literature a multiplicity of variants of

generalized cross-correlation have been presented, basically introducing a weighting factor in order to take into account the statistics of source signal and noise. If a normalization factor is applied in order to preserve only the phase information:

$$G_{PH,m}(f) = \frac{S_m(f)S_r(f)^*}{\|S_m(f)\|\|S_r(f)\|} \quad (4)$$

the Cross-Power Spectrum Phase (CSP) [6] or Phase Transform (PHAT) [1] is obtained as:

$$CSP_m(t) = IFFT[G_{PH,m}(f)] \quad (5)$$

Considering that the delay in time domain corresponds to a phase rotation in frequency domain, it turns out that the IFFT of the function (4) presents a delta pulse centered on the delay τ. The delay estimate is derived from:

$$\tilde{\tau}_m = \arg\max_t CSP_m(t) \quad (6)$$

Thus, the information in the CSP peaks, where the inter-channel coherence is higher, locates the delays, and indirectly the source position via trigonometry: the CSP can drive a D&S beamformer (CSP-D&S) toward the maximum coherence directions. As we will show, these directions are sometimes the main reflections rather than the direct path from the source to the array and this does generally not imply to have a higher recognition rate especially if sound sources are not facing the microphone array. We will also show that, though the theta-D&S is useful to evaluate the best directions in the pi-space for recognition, a CSP-D&S works generally better.

5 The Limabeam algorithm

Once the τ_m have been calculated, either by fixing a certain angle or by performing TDE via CSP, one can further shape the directivity pattern by finding the optimal weights $W_m(f)$ in Equation (1). These filters can be fixed or adapted on a per-channel or per-frame basis, depending on a chosen criterion. In this work we seek to find optimal filters which increase the recognition performances rather than the Signal to Noise Ratio (SNR): the goal is reach by using the Limabeam algorithm. Indeed, this algorithm, introduced by Seltzer [7, 4], estimates an adaptive filter-and-sum beamformer. In the discrete time domain Equation 2 becomes:

$$x[k] = \sum_{m=1}^{M} h_m[k] * s_m[k - \tau_m] \quad (7)$$

where $h_m[k] = IFFT(W_m(f), k)$ is the FIR filter for the m-th channel, $*$ denotes convolution and k is the time index. The whole set of FIR coefficients of all microphones can be represented by a super-vector $\mathbf{h}$. For each frame, recognition features can be derived and expressed in function of $\mathbf{h}$:

$$\mathbf{y}_L(\mathbf{h}) = \log_{10}\left(W \,\|\,\mathrm{FFT}(\mathbf{x}(\mathbf{h}))\|^2\right) \quad (8)$$

where $\mathbf{x}(\mathbf{h})$ is the observed vector, $\|\,\mathrm{FFT}(\mathbf{x}(\mathbf{h}))\|^2$ is the vector of individual power spectrum components, W is the Mel filter matrix and $\mathbf{y}_L(\mathbf{h})$ is the vector of the Log Filter Bank Energies (LFBE). Cepstral coefficients are derived via a DCT transform:

$$\mathbf{y}_C(\mathbf{h}) = \mathrm{DCT}\left(\mathbf{y}_L(\mathbf{h})\right). \quad (9)$$

Limabeam aims at deriving a set of *M* FIR filters, which maximize the likelihood of $\mathbf{y}_L(\mathbf{h})$ given an estimated state sequence of a hypothesized transcription. This is expressed by:

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}} P(\mathbf{y}_L(\mathbf{h})\,|w\,) \quad (10)$$

where w is the hypothesized transcription, $P(\mathbf{y}(\mathbf{h})\,|w\,)$ is the likelihood of the observed features given the transcription considered and $\hat{\mathbf{h}}$ is the FIR parameter super-vector derived. The optimization is done via the non-linear Conjugate Gradient. The state sequence can be estimated either using the array beamformed output (Unsupervised Limabeam or UL) or, alternatively, assuming that the correct transcription is available (Oracle Limabeam or OL). In both cases the filters are estimated on-line, meaning that for each test sentence a new set of filters is generated starting from the D&S configuration. Alternatively, one can optimize just one set of filters and keeping it for the whole session (Calibrated Limabeam or CL). More details can be found in [8].

6 Improving Limabeam: Nbest and TCL

In our previous work [9] we showed that both in simulations and in a real environment affected mainly by additive noise, Limabeam can be improved. This is done by optimizing in parallel the multi-channel signal not just on the first hypothesized transcription, but on the *N*-best hypotheses, where *N* is as high as possible. The criterion adopted is

$$\hat{\mathbf{h}}_n = \arg\max_{\mathbf{h}} P(\mathbf{y}_L(\mathbf{h})\,|w_n\,) \quad (11)$$

where w_n is the *n*-th hypothesized transcription at first recognition step, $P(\mathbf{y}(\mathbf{h})\,|w_n\,)$ is the likelihood of the observed features given the *n*-best transcription considered. Note that Equation (11) is equivalent to Unsupervised Limabeam when *n* is 1. After all the *N*-best FIR vectors are optimized in parallel, new features are calculated and recognition is performed. The transcription which gives the ML is then chosen:

$$\hat{n} = \arg\max_n P(\mathbf{y}_C(\hat{\mathbf{h}}_n)\,|\hat{w}_n\,) \quad (12)$$

where $\hat{w}_n$ is the transcription generated at second step recognition and $\hat{n}$ is the index of the most likely transcription, which is $\hat{w}_{\hat{n}}$. The proposed *N*-best approach improves the performances of the Unsupervised Limabeam. In this work we propose to improve also the Calibrated Limabeam by estimating the filters differently. Instead of calibrating the set of filters on a sentence extracted from the test set, we try to derive a set of filters which improves performances independently on the position of the speaker. To this aim, we optimize filters using clean speech from the Training set convolved with a set of room impulse responses which do not match the test conditions. We find that for sufficiently short FIR filters, the recognition performances is independent on the set of room impulse responses used for performing the proposed Training-set Calibrated Limabeam (TCL). Our experiments will show that, when no information about the speaker location is available, TCL performs on average better then any version of the Limabeam algorithm.

7 Matched Filtering

The techniques presented so far do not make use of any knowledge of the speaker position in the room. Being that available, one can use the punctual information related to a specific pair "source-microphone" for generating the so-called *Matched Filter*, that realigns not only the primary delay (usually associated to the direct path) but also the secondary delays. In short, the filter is derived from a flipped and truncated version of the impulse response [10]. If h_m^p, impulse response in position p with respect to microphone m is known, the following filters are considered:

$$g_m^p[k] = h_m^p[K-k] \tag{13}$$

where K in the final filter length. The enhanced signal is the product of the consequent "filter-and-sum" processing:

$$x[k] = \sum_{m=1}^{M} g_m^p[k] * s_m[k] \tag{14}$$

which is equivalent at shaping the directivity pattern in Equation (1) with $W_m(f) = FFT(g_m^p[k])$. Having knowledge of the impulse responses at test time can provide an upper bound for performances. We propose to get a potentially higher upper bound if MF is used instead of D&S prior to Limabeam. In this case Equation (14) becomes:

$$x[k] = \sum_{m=1}^{M} h'_m[k] * s_m[k] \tag{15}$$

where $h'_m[k] = h_m[k] * g_m^p[k]$ is the per-channel filter to be optimized.

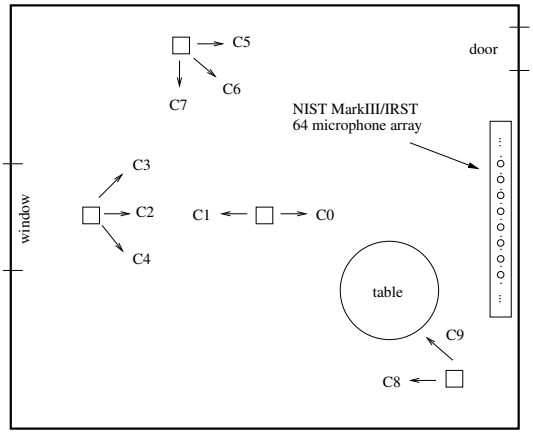

Figure 2. *Map of the ITC-irst CHIL room (6m × 5m), reporting on positions of array and acoustic sources.*

8 Experiments and Results

The experimental setup consists of a recognition task of 1001 connected English digits sentences: the original TI-digits signals have been reproduced by a high-quality loudspeaker in the CHIL room available at ITC-irst (T_{60} is approximately 0.7 s) and acquired at a sampling frequency of 44.1kHz by means of a linear array of 64 microphones (Mark III board). This test set has been evenly divided in subsets, varying position and orientation of the loudspeaker with respect to the array for a total number of 10 different configurations. Figure 2 identifies in the room map the 10 subsets, indexed by C0 to C9. As a result the Signal-to-Noise-Ratio, evaluated at one microphone of the array, varies from 10 to 25dB, depending on position, orientation and energy of the original signal.

Experiments were conducted using the HTK HMM-based recognizer [11] trained on the clean TI-digits corpus. Word models are represented by 18 state left-to-right HMMs. Output distributions are defined by 1 Gaussian pdf. The training set consists of 8440 utterances, pronounced by 110 speakers (55 men and 55 women). The FIR filters to be optimized by the Limabeam are 10 taps long. The feature extraction in the front-end of the speech recognizer involves 12 Mel Frequency Cepstral Coefficients and the log-Energy together with their first and second derivatives, for a total of 39 coefficients. Features were calculated every 10 ms, using a 25 ms sliding Hamming window. The frequency range spanned by the Mel-Scale filterbank was limited to 100-7500 Hz to avoid frequency regions with no useful signal energy. Cepstral Mean Normalization is applied. A sub-array of the MarkIII was chosen for our experiments: we used 8 microphones spaced by 4 cm. This was done both to

get a high directivity under spatial aliasing constraints and to limit the system complexity (the more the microphones, the higher the number R of beams of Equation (3) and the more difficult the filter optimization).

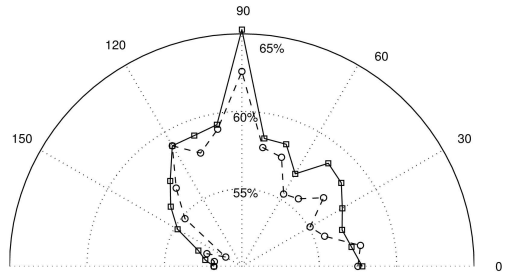

Figure 3. *Polar Recognition Directivity Pattern when speaker is in configuration C2: the array points with a very narrow beam toward the speaker, while smaller sidelobes between 0° and 60° collect minor reflections. Unsupervised Limabeam (solid line) almost always gains on theta-D&S (dashed line). The pattern magnitude is measured in WRR, starting from 50%.*

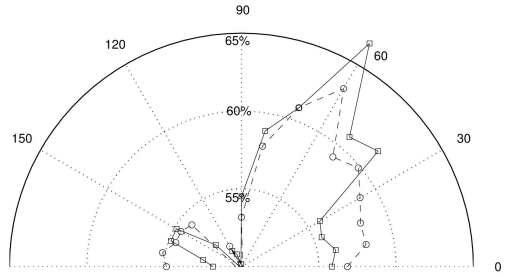

Figure 4. *Polar RDP when speaker is in configuration C5. The array definitely points toward the speaker, which in turns faces the door. Early reflections on the closer side wall are beneficial between 30° and 60°. UL is very effective in the most relevant direction.*

Figures 3, 4, 5, and 6 show that in all scenarios the RDP has a main lobe corresponding to the speaker direction, i.e. the direct path. Both C2 and C5 represent favorable cases, where the speaker (located at 90° and 60° respectively) is pointing to the array, while in C7 and C9 the direct path reaches a wall first, but the RDP points to the speaker anyway (located at 60° and 150° respectively). This is evidenced by the presence of larger recognition sidelobes. We verified that peaks of the RDP (e.g., in C9) can correspond to the main reflections detected by the CSP: Figure 7 reports the superposition of a CSP and a RDP in Cartesian Coordinates. The configuration C7 is the most difficult and challenging to evaluate. In this case a beamformer is effective only if it points at a specific direction in the space. In particular here a theta-D&S performs better than a CSP-D&S, because the former directs the beamformer to the weak-coherence path, which is more relevant from a recognition oriented perspective than the strong, main reflection. We tried to manage this situation by automatically selecting the two main peaks, sentence by sentence (this was done simply by finding the maxima of the CSP function with linear regression and zero crossing of the first CSP discrete derivative) and we achieved the single-channel performances, which is roughly the average of the two main peaks performances. In all the scenarios depicted the Unsupervised Limabeam is effective, and the best relative improvements over D&S are obtained if it is applied to directions

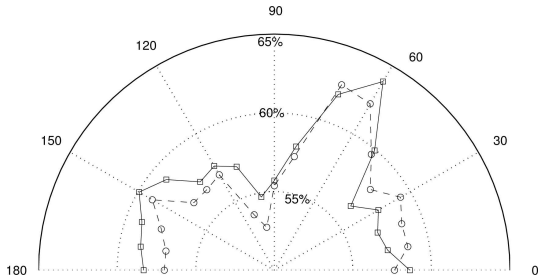

Figure 5. *Polar RDP when speaker is in configuration C7: the array points toward the source, located at 60°, but a large lobe 'seeks' the main reflection at 150°. In this configuration the CSP-D&S points to the latter recognition lobe, which is related to a CSP peak with more coherence but less impact on recognition performances. UL gains over theta-D&S from 45° to 180°*

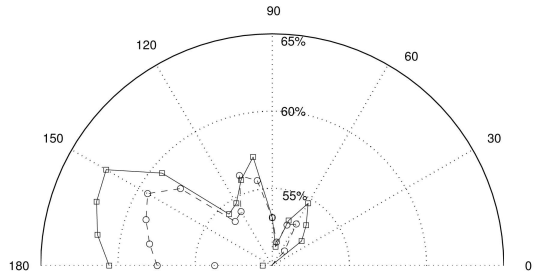

Figure 6. *Polar RDP when speaker is in configuration C9: the array points at the speaker, but two lobes collect the contribution of the correspondent main reflections. UL is always effective.*

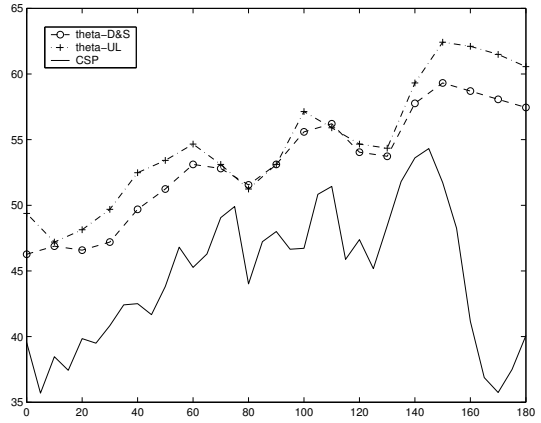

Figure 7. *RDP in Cartesian Coordinates for configuration C9: the RDP peaks are well related to the main CSP peaks. CSP peak heights were normalized for plotting purposes only.*

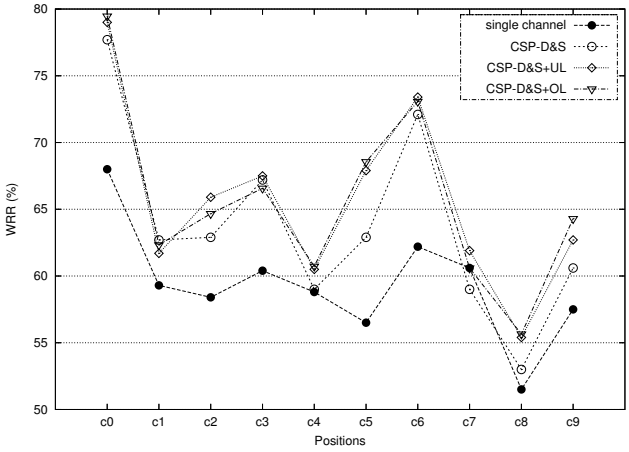

Figure 8. *Baseline results: Word Recognition Rates (%) in the 10 test position using single-channel, Delay-and-Sum beamforming and Unsupervised Limabeam.*

toward the speaker. Apart from C7, the CSP-D&S performs generally better and its filters can be used to initialize any Limabeam-based algorithm. Figure 8 shows the WRR in function of the 10 test positions: clearly the use of D&S is improving performances and as much as the speaker is both pointing to and close to the array. This is intuitive, because by pointing to the speaker, performances tend to be proportional to the the signal-to-reflection ratio. UL and OL both give improvements on average over D&S. Table 1 shows the results relative to CSP-D&S and its coupling with UL, i.e. when estimation on both the delays and the filters is done without any prior knowledge of the environment.

	single mic.	CSP-D&S	CSP-D&S+ UL
ave(c0-c9)	59.3	63.7	65.6

Table 1. *Table reporting Average Word Recognition Rates (%) over the 10 test configurations.*

TCL is a version of the Calibrated Limabeam where filters are estimated offline on a contaminated training phrase: it differs from CL because the contamination is done with impulse responses of positions *different* from the one in test set. The filter length has been limited to 10 taps because we verified that any technique based on Limabeam (with 8 microphones) improves performances up to a certain filter length: Figure 9 reports the WRR of a TCL in function of the number of taps. Note that performances of TCL for position cX are measured as an average of the performances when the training impulse response owns to all the positions except cX. Indeed, Figure 10 shows that training with the (very different) impulse responses from positions c0, c1, or c8, lead almost to the same results. In this sense TCL provides a sort of "room equalization", because it can estimate filters that perform in the same way across all the positions and thus are independent from them. Furthermore we compare all the Limabeam-based techniques in Figure 11 position by position and in Table 2 on average. The N-best approach was successfully tested in another environment and with mostly additive noise [9]. We observe that in such a reverberant environment a technique based on calibration is more suitable than a sentence-by-sentence adaptation: in fact the filters generated, for example, by the UL, are very similar across the sentences, and being limited by few taps increases the likelihood of the different positions.

	UL	OL	CL	Nbest	TCL
ave(c0-c9)	65.6	65.6	67.3	66.5	67.9

Table 2. *Table reporting Average WRR (%) over the 10 test configurations for Limabeam-based algorithms.*

This is why TCL has the highest performances in six positions out of ten, while the other four it is second only to CL. Being the filters so short, the effectiveness of the Limabeam-based techniques resides in modifying just the spectral tilt: this motivates us in searching for a possible longer filter, which could represent an upper bound for our performances.

We found this filter being the Matched Filter: Figure 12 shows the WRR in function of the MF length: depending on the position, the peak in accuracy is reached for different lengths. However, this length may well correlate with

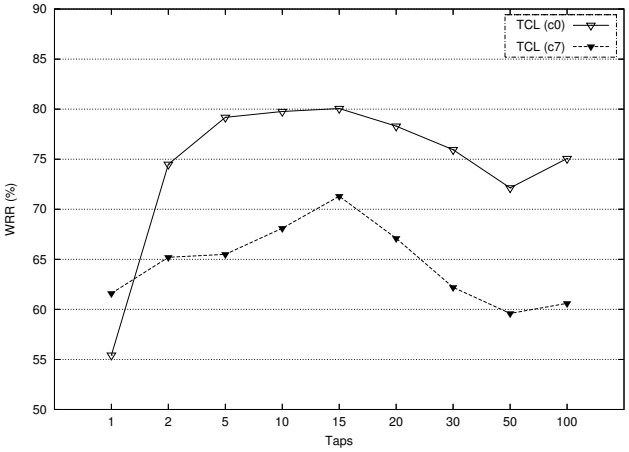

Figure 9. *WRR for TCL technique as a function of the filter's length.*

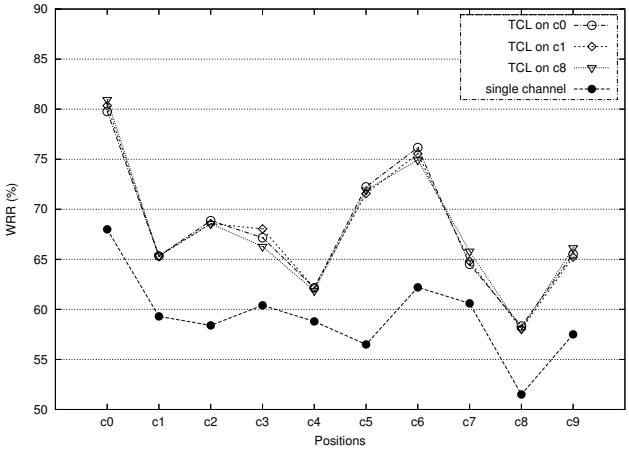

Figure 10. *WRR (%) in the 10 test position for TCL filter estimated on c0, c1, and c8 position.*

the relative T60: the MF is effective once it includes the direct path and the main reflections. In c0 these reflections are 1000 taps away from the direct path and the accuracy curve slowly lowers down, while for c1 the optimal length is around 3000 and for c8 8000, which means there are useful (from a recognition point of view) reflections at about 70 and 180 ms respectively from the direct path.

Table 3 reports on the average performances across positions of the MF (limited to 1500 taps for every position), also coupled with OL, the latter meaning that full knowledge of the target is given (i.e. the exact impulse response and the correct sentence for optimization). Results with MF are high compared to Table 2 (35% relative improvement of MF+OL over CSP-D&S), which means that there is still a high margin for technique aiming at finding a maximal WRR set of filter for a multi-channel signal. It is also worth noting that the relative improvement of the OL after MF is used is 12.5%, while after CSP-D&S is used is 5.2%, showing that the initialization of filters is crucial for a Limabeam-based technique. The Table also reports on the *UnMatched Filtering*, which corresponds to applying Matched filters of positions cX to test position cY, exactly as we did with TCL: it is worth noting that with MF performances drop down dramatically if the position impulse response is not matched, thus this a-priori knowledge is not interchangeable between test sets, as it happened with TCL. MF is thus very effective but not realistic for a real world application.

	single mic.	unMF	MF	MF+OL
ave(c0-c9)	59.3	53.2	73.0	76.4

Table 3. *Table reporting Average WRR (%) over the 10 test configurations for MF-based algorithms. Matched Filtering (MF) requires additional knowledge (i.e., room impulse response) but provides a tangible performances boost. The adoption of unMatched filters (unMF), on the other hand, is harmful.*

9 Discussion and Conclusions

In this work we have investigated the use of microphone array processing in a real reverberant room, analyzing the impact of different beamforming techniques on performances measured in terms of Word Recognition Rate on a digit recognition task. Several beamforming techiques based on inter-channel delay handling (theta-D&S, CSP-D&S) and on a likelihood-based filter-and-sum beamformer (UL, OL, CL, N-Best UL, TCL) were presented and tested, showing that, in some configurations, critical aspects can be the correct estimation of the inter-channel delay and the initialization of the filters. Performances are relatively high when the speech source is directed to the sensors as well as the array is steered toward the source, but in this case it is very sensitive to steering errors. To cope with these errors, a CSP-driven beamformer can automatically locate the useful wavefront. On the other hand, when sources and microphones are not faced to each other, which mimic for example differently head oriented speakers, there is a direct correspondence between the peaks of the CSP and RDP figures. A possible relation between their relative magnitude is under investigation. Future work will be directed to establish a criterion for selecting higher recognition lobes independently of speaker location and orientation. Furthermore, in all the scenarios we were able to get further improvements, with respect to both a theta-driven and CSP-driven D&S, by using the Unsupervised Limabeam, which

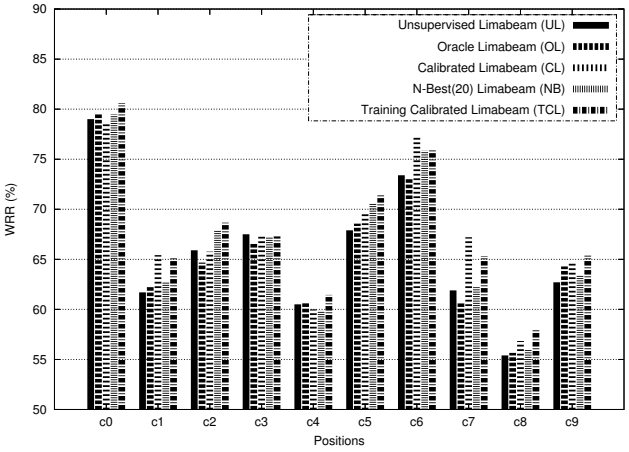

Figure 11. *WRR (%) adopting Limabeam-based techniques.*

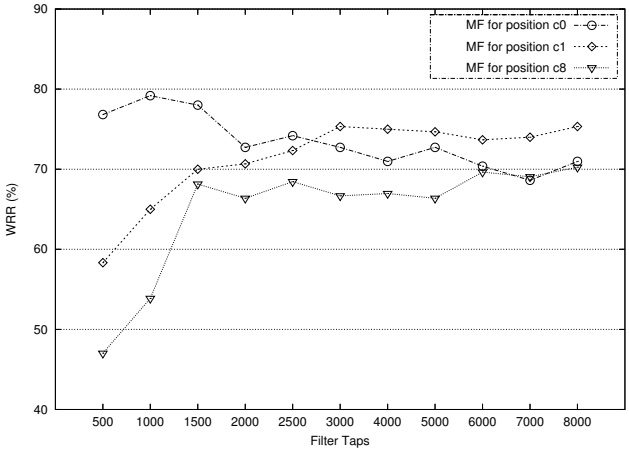

Figure 12. *WRR (%) for Matched Filtering as a function of number of taps.*

is as much effective as the initial configuration of FIR filters steers the array to direction corresponding to high recognition lobes. The most performant version of Limabeam is the proposed TCL, which derives a set of calibration filters from a clean speech sentence contaminated with impulse responses which do not match the test conditions. However, the improvement is limited and the few number of taps used do not allow to consider the main reflections at the current sampling rate. The use of Matched Filtering, which well couples with Limabeam but can't be used in practice, shows that there exist a set of (long) filters which dramatically increase performances, that the working point in which the optimizations starts is crucial and that the margin of improvement is still high for technique aiming at finding a filter optimum from the recognition point of view. A method which can automatically select, based on different confidence measures, the correct Matched Filter sentence by sentence is under investigation.

References

[1] M. Brandstein and D. Ward, *Microphone arrays - signal processing techniques and applications*, New York: Springer-Verlag, 2001.

[2] L. Griffith and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," in *IEEE Trans. on Antennas and Propagation*, 1982, vol. AP-30, pp. 27–34.

[3] O. Frost, "An algorithm for linearly constrained adaptive array processing," in *Proceedings of the IEEE*, 1972, vol. 60, pp. 926–935.

[4] M. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," in *IEEE Trans. on Speech and Audio Procesing*, September 2004, vol. 12(5), pp. 489–498.

[5] D. Johnson and D. Dudgeon, *Array signal processing*, Prentice Hall, 1993.

[6] M. Omologo and P. Svaizer, "Acoustic event localization using a cross-power spectrum phase based technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1994.

[7] M. Seltzer and B. Raj, "Speech recognizer-based filter optimization for microphone array processing," in *IEEE Signal Processing Letters*, March 2003, vol. 10(3), pp. 69–71.

[8] M. Seltzer, *Microphone array processing for robust speech recognition*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2003.

[9] L. Brayda, C. Wellekens, and M. Omologo, "Improving robustness of a likelihood-based beamformer in a real environment for automatic specch recognition," in *Proceedings of Specom*, St.Petersbourg, Russia, 2006.

[10] Flanagan J.L., Surendran A.C., and Jan E.E., "Spatially selective sound capture for speech and audio processing," *Speech Communication*, 1993.

[11] S. Young and et al., *The HTK Book Version 3.0.*, Cambridge University, 2000.