ROBUST SPEECH RECOGNITION WITH MICROPHONE ARRAYS IN MULTI-ROOM HOME ENVIRONMENTS

José Matos\textsuperscript{1,2}, Alberto Abad\textsuperscript{1,2}, Isabel Trancoso\textsuperscript{1,2}

\textsuperscript{1} Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal ; \textsuperscript{2} INESC-ID, Lisbon, Portugal

ABSTRACT

This paper presents a set of exploratory experiments addressed to analyse and evaluate the performance of baseline speech processing components for distant voice command recognition applications in domestic environments. The analysis, conducted in a multi-channel multi-room scenario, showed the importance of adequate room detection and channel selection strategies to obtain acceptable performances.

Three different computationally inexpensive channel selection measures for room detection, channel selection and cluster selection have been investigated. Experimental results show that the strategies based on envelope-variance measure consistently outperformed the remaining methods investigated, and particularly, that channel selection strategies can be more convenient than baseline beamforming methods, such as delay-and-sum, for this type of multi-room scenarios.

Finally, the problem of speech/non-speech detection was investigated in a multi-microphone multi-room environment, exploring different approaches that take advantage of the availability of multiple channels. Combination of microphone selection strategies with speech/non-speech detection was also investigated for simultaneous detection and localization of speech events inside a room.

This paper is a summary of the best techniques analysed in the thesis.

Index Terms— distant speech recognition, multi-microphone processing, beamforming, microphone selection, home control applications

1. INTRODUCTION

The DIRHA project\textsuperscript{1} addresses the challenge of distant-speech recognition in a home environment, with a very realistic and complex application scenario, using a microphone network distributed over the different rooms of an apartment. The system is “always-listening”, and an important challenge is to develop a solution that reduces false alarms due to misinterpretation of human-human conversations and other non-speech sounds. Speech enhancement and recognition methods for this type of application have been widely described in the literature, but they do not address the realistic scenario of sound sources occurring in different rooms. The focus of the paper is on multi-room, multi-microphone solutions adequate for the home scenario. The paper starts by describing the data collection process, explaining how simulated corpora were obtained for different languages. The bulk of the paper is devoted to the description of the different experiments on the task of recognizing read commands.

2. CORPORA

This section describes the corpora used to build and evaluate the approaches studied in this work. The database BD-PUBLICO is a corpus used for training the models of the various recognizers. For testing purposes, we use the DIRHA-EP SimCorpus that accurately represents the acoustic characteristics of a domestic environment.

2.1. The BD-PUBLICO Corpus

The BD-PUBLICO database (Base de Dados em Português Europeu, vocabulário Largo, Independente do orador e fala Continua) is aimed at the development of large vocabulary, speaker-independent continuous speech recognition systems in European Portuguese. The text material for the read sentences are extracted from the Portuguese newspaper PÚBLICO, consisting of 6 months of news, totaling 10M words and 156k different forms. The corpus is split into 3 sets:

- Training set: 80 sentences plus 3 calibration sentences for each speaker.
- Development set: 40 sentences plus 15 speaker-adaptation sentences per speaker.
- Evaluation set: 40 sentences plus 15 speaker-adaptation sentences and 3 calibration sentences for each speaker.

Speaker selection was done among undergraduate and graduate students from IST. Ages ranged between 19 and 28 and a broad coverage of accents was obtained. The corpus contains a total of 120 speakers recordings: 100 for the training set (50 male and 50 female); and 20 speakers (10 male and 10 female) divided equally in the 5K word sets (evaluation / development). Each recording session resulted in approximately 15 minutes of speech. The recordings were done in a sound proof room using a high quality microphone at 16kHz sampling frequency.

For a more detailed overview of the database structure and contents, one can refer to [2].

2.2. The DIRHA SimCorpus for European Portuguese

The DIRHA SimCorpus is a multi-microphone and multi-language database containing simulated acoustic sequences derived from the microphone-equipped apartment located in Trento (Italy), named ITEA, depicted in Figure 1. The simulated corpora for the different languages -including European Portuguese (EP) - were produced

\textsuperscript{1}This work was partially funded by the European Union, under grant agreement FP7-288121. The authors would like to thank their colleagues in the DIRHA consortium.

\textsuperscript{1}http://dirha.fbk.eu
thanks to a technique that reconstructs, in a realistic manner, multi-

microphone front-end observations of typical scenes occurring in
a domestic environment. For each language, the corpus contains a
set of acoustic sequences of duration 60 seconds, at 48kHz sam-
pling frequency and 16-bit accuracy, observed by 40 microphone
channels distributed over five rooms. Each sequence consists of
real background noise with superimposed localized acoustic events,
occurring randomly (and rather uniformly) in time and in space
(within predefined positions) with various dynamics. The acoustic
wave propagation from the sound source to each single microphone
is simulated by convoluting the clean signals with the respective
measured impulse response (IR). Acoustic events are divided into
two main categories, i.e., speech and non-speech. Speech events in-
clude different types of utterances (i.e., phonetically-rich sentences,
read and spontaneous commands, conversational speech). Non-
speech events have been selected from a collection of high-quality
sounds typically occurring within a home environment (e.g., radio,
TV, appliances, knocking, ringing, etc.).

For the DIRHA SimCorpus in EP, hereinafter referred to as
DIRHA-EP SimCorpus, a clean-speech data set of very high quality
close-talking speech signals was collected to derive the simulated
corpus. The data set contains 20 speakers with an equal gender
distribution, ageing between 25 and 50. The EP simulated corpus
is divided into two chunks (dev and test) containing 75 acoustic
sequences each, with 10 different speakers in each data set [11].

3. BASELINE FOR DISTANT SPEECH RECOGNITION IN
PORTUGUESE

The baseline ASR system for EP in DIRHA has been developed
and assessed with the HTK toolkit [1], using the BD-PUBLICO

corpus. A bi-gram language model trained with 11M words is provided
with this corpus. The selected closed-set vocabulary includes 6,618
unique words. The features used to represent acoustic information
are the traditional 13-dimensional MFCCs, augmented by their first
and second derivatives, and mean normalized, thus reaching a di-
mensionality of 39. Following acoustic feature extraction, HMM
training was carried out using the typical HTK pipeline toolkit [1].
First, a set of monophone models was trained using “flat-start” ini-
itialization (39 phone classes). Then, training of monophones, tri-
phones, state tying, and Gaussian mixture splitting was performed
until final cross-word tied-state context-dependent triphones of 3
states and 16 Gaussians per state were built. In the state tying pro-
cess, a decision tree clustering strategy was applied, using a set of
phonetic questions adequate for EP.

A data-simulation based approach was adopted to evaluate ASR
in the DIRHA far-field environment, by artificially convolving the
“clean” recordings with IRs measured at a number of locations in
the apartment, and contaminating them with noise at various SNR
values. This resulted in two sets of simulated far-field data, one us-
ing a more controlled contamination approach (“reverb1”), and
another using a wider contamination parameter variability applied in
a random fashion (“reverbR”) [4]. The dev set was used for ad-
justing the decoding parameters (in clean matched conditions) which
henceforth were kept constant for all experiments. Unfortunately,
there is no warranty that the large acoustic variability of the DIRHA
multi-room and multi-channel scenario is well-represented by only
one of these single environmental conditions. In order to partially
mitigate this problem, we followed a quite straightforward approach
that consists of using all the training data of the three conditions
together to train a single multi-condition acoustic model. This ap-
proach resulted in a better performance than the single condition
models (see [12] for details), yielding WERs of 30.04% and 32.67%
for the dev and test “reverb1” sets, and 33.07% and 34.53%
for the dev and test “reverbR” sets, respectively. The multi-
condition models were therefore used in all remaining experiments.

4. RECOGNITION OF VOICE COMMANDS IN
MULTI-ROOM SCENARIOS

This section presents the first exploratory studies conducted on the
DIRHA-EP multi-room and multi-channel corpus. The experiments
are focused on the task of recognizing read spoken commands. For
this purpose, the spoken commands are extracted from the two sets
(dev and test) of the 1 minute DIRHA-EP SimCorpus simulations

using the voice activity ground truth information. There are a total
of 75 simulations per dataset and each simulation contains one read
command, which results in a total of 150 extracted speech segments
that are used for evaluation. For the command recognition task, an
equally-likely finite state grammar formed by all the unique possible
command sentences is used.

4.1. Analysis of performance in multi-room environments

In contrast to other existing corpora commonly used in the field of
robust and far-field speech recognition research, the acoustic events
of interest in DIRHA do not always happen in the same room where
all the microphones are placed. Instead, acoustic events (speech or
any other) may happen anywhere inside the ITEA apartment and
they are simultaneously collected by a network of 40 different micro-
phones that are also distributed in the different rooms of the house-
hold. Consequently, the challenges faced at the DIRHA project

go far beyond the classical problem of far-field speech recognition.
Thus, it became necessary to understand how critical is the selection
of a specific microphone or group of microphones in multi-room
speech recognition tasks.

Figure 2 shows a pictorial summary of the recognition perfor-

mance for each one of the microphone channels on the dev set. The
results are the error rates obtained by each channel when recognizing
only the speech events taking place in a specific room. The results
for all the 75 events are also provided in the first row. The darker
(red) each cell is, the larger the corresponding WER is. Notice that
the microphone channels have been sorted per room (the first char-
acter in the name of each microphone identifies the room). Thus, the
pseudo-diagonals of the coloured matrices represent matched condi-
tions in which the speech events occur in the room where the micro-
phone is located. Notice also that the number of spoken commands
is not the same in all rooms. This number is provided in the y-axis
together with the name of the room. Consequently, the microphones
located in rooms where spoken commands are more frequent obtain
better overall performances as it could be expected. One can eas-
ily identify the room matched conditions, since they are consid-
erably lighter than the regions outside the pseudo-diagonal. Although
there are differences among channels inside the same room, these
differences are considerably smaller than the ones observed when
compared with the performance of a microphone outside the room.
Hence, as it could be expected, the room information is fundamental.

The first row of Table 1 shows the WER results computed in the
dev and test datasets using the microphone that obtained the best
overall performance in the dev set, that was LA5. This living-room
microphone was in practice also the best overall performing channel
in test data. Contrarily to typical results, these experiments have
shown that the test set seems considerably easier than dev. The
second row of Table 1 shows the performance of a system that knows
Fig. 1. Outline of the microphone set-up. Black dots represent microphones, while boxes and arrows represent available positions and orientations.

<table>
<thead>
<tr>
<th>Microphone selection</th>
<th>testing conditions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dev</td>
<td>test</td>
</tr>
<tr>
<td>overall best-mic (LA5)</td>
<td>35.28</td>
<td>25.62</td>
</tr>
<tr>
<td>room-aware random-mic</td>
<td>11.99</td>
<td>9.64</td>
</tr>
<tr>
<td>room-aware best-mic</td>
<td>6.74</td>
<td>8.94 (4.47)</td>
</tr>
<tr>
<td>oracle mic per-event</td>
<td>4.94</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 1. Average WER (%) performance using different microphone channel selection strategies exploiting knowledge about the room where an event occurs.

The room where an event happens and that randomly selects a microphone of the given room to recognize that speech event. This score has been obtained computing the mean performance of 500 different random selections (the standard deviation was ∼ 2 – 3%). The performance of this “room-aware” approach is considerably better than the one obtained by the best overall microphone. In other words, if the room in which an event occurs is a given information, a simple straightforward approach for improved recognition in the DIRHA scenario consists of selecting one of the microphones of that room to process the command. The third row of Table 1 reports the performance of a system that knows the room where an event occurs and that recognizes this event using the microphone that obtained the lowest average error for the events of that room in the dev dataset. While the random microphone selection is clearly outperformed by this oracle approach in dev, the difference in test is not so meaningful. In other words, the best microphone is largely dependent on the specific event and the best average configuration in dev does not necessarily provide better results in test than a simple random selection approach. For comparison purposes, we report in brackets the performance that is obtained in the test set. The considerable improvement reinforces the importance of a microphone selection strategy that depends on the specific event rather than a general per room configuration. Finally, the upper-bound performance that could be attained with a perfect microphone selection algorithm is reported in the last row of Table 1. This perfect oracle was computed by simply selecting the channel that provides the lowest error for any specific speech event. This ideal figure represents the potential of channel selection algorithms, even in the case of extremely challenging multi-microphone and multi-room acoustic environments.

4.2. Beamforming with multiple microphone clusters

The delay and sum beamformer [6] is one of the simplest and most efficient microphone array approaches. It consists of the alignment of the microphone signals to compensate for the different acoustic path lengths from the source to the microphones, followed by the addition of the time-aligned signals. For a multi-channel system with $M$ microphones, this can be expressed as

$$y(t) = \sum_{m=1}^{M} \alpha_m x_m(t - \tau_m)$$

(1)

where $x_m(t)$ are the different microphone channels, $\alpha_m$ are the channel gains and $\tau_m$ are the time delays of arrival. Typically, one of the microphones is used as the reference channel, and the remaining microphones are compensated with respect to the reference. Simplicity is the most important strength of this approach, making it a practical choice for many microphone array applications. But such a simple spatial filter can only partially suppress directional interference. Given that there is a large amount of microphones in our scenario, our hypothesis is that applying multi-channel processing to specific sub-sets of microphones can be of more benefit than for instance applying beamforming to all the available microphones in a room. In these experiments, we assume that we know the position of the speech source, so the propagation delays can be compensated for all the microphones. Moreover, we consider an equal gain $\alpha_m = \frac{1}{M}$ for all the channels involved in a delay and sum cluster.
Figure 3 shows a pictorial summary of the recognition performance obtained with delay and sum beamformer configurations for the dev set. We have considered 18 delay and sum beamformer configurations: 14 formed by the different microphone clusters of the apartment; 3 beamformers processing all the microphones of the Living-room, Bedroom and Kitchen, respectively; and 1 beamformer that processes the 40 channels. Like in the previous single channel pictures, the cells outside the pseudo-diagonal represent the mismatched scenario in which a beamformer of a different room is selected to process the events of a particular room. Notice that, in contrast to the single channel analysis, this is an awkward scenario, since we are assuming that we have perfect knowledge about the source position to correctly steer the beamformers. Comparing with the previous picture, the colours are “lighter”, that is, the overall performance is better when using beamforming signals.

Although in the case of beamforming experiments, the valuable comparisons are between different cluster configurations inside the same room, results obtained with the best overall beamformer when processing all the events are also reported in the first row of the summary Table 2. Notice the significant improvement with respect to LA5 results of Table 1. Table 2 also summarizes some of the multi-channel “room-aware” speech recognition configurations performance. The second row shows the performance obtained when a delay and sum beamformer of all the microphones of a room is used to spatially filter and later recognize the events occurring in that specific room. From rows 3 to 5, similarly to the previous section 4.1, we report respectively the performance obtained when a random beamformer of the correct room and the best average beamformer of the correct room are selected to process and recognize the speech commands. Notice that the simple delay and sum approach with all microphones achieves excellent results, close to the best cluster selection in dev, and equalling its performance in the test set. Comparing the “room-aware” delay and sum approaches against the corresponding single channel performances, it is worth noting that there is not a very significant difference if the best cluster or channel is selected. The largest difference is observed mainly in the case of “random” selection. In other words, although the use of delay and sum beamformers in combination with cluster selection provides reduced performance improvements in ideal cluster/microphone selection conditions, they are definitely more robust to erroneous channel selections. Finally, the last row of Table 2 reports again the oracle results in which the appropriate delay and sum cluster is selected per-event. Notice that these results are slightly worse than the ones obtained without beamforming. This may happen because an event that was correctly detected using the channel selection approach is no longer correctly recognized using the beamformed signals.

### 4.3. Automatic channel selection and room detection

The two previous sections showed the importance of room knowledge and also the potential of adequate microphone(s) selection strategies. In this section, we explore two simple low-cost methods for automatically selecting either the room, the beamforming cluster or the microphone channel. To the best of our knowledge, this is the first reported work concerning channel selection strategies for multi-room environments.

**Table 2.** Average WER (%) performance using different cluster selection strategies for beamforming exploiting knowledge about the room where an event occurs.

<table>
<thead>
<tr>
<th>Cluster selection</th>
<th>testing conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dev</td>
</tr>
<tr>
<td>overall best-beamf (beamf_all)</td>
<td>17.98</td>
</tr>
<tr>
<td>room-aware allmics_room-beamf</td>
<td>8.54</td>
</tr>
<tr>
<td>room-aware random-beamf</td>
<td>10.08</td>
</tr>
<tr>
<td>room-aware best-beamf</td>
<td>6.74</td>
</tr>
<tr>
<td>oracle beamf per-event</td>
<td>5.17</td>
</tr>
</tbody>
</table>
4.3.1. Channel selection (CS) approaches

An overview of some of the most remarkable methods proposed for the microphone selection problem for speech recognition in multi-channel single-room environments can be found in [9]. In practice, the most common approaches are either related to some direct measurement of the signals (SNRs, distortion, correlation, etc.), some indirect relation (knowledge of the position and orientation, knowledge of the room impulse response, etc.) or obtained based on the multi-channel recognition results (likelihood hypothesis, hypothesis combination, etc.). The latter, which are usually named back-end approaches, are able to obtain very competitive performances but they present some limitations: they are computationally inefficient, since they usually need to perform recognition in all channels, and they may suffer from normalization problems due to different acoustic realizations[8]. These problems are in fact exacerbated when many microphones from different rooms are available, like in the case of DIRHA. Consequently, back-end methods are not a convenient solution for our particular task. On the other hand, methods based on indirect relations like position or room impulse responses may be impractical in real world applications due to the difficulty to estimate this type of information. Thus, the focus of this set of exploratory experiments is mainly concentrated in direct measurement based approaches for microphone selection. Particularly, we have considered two alternative signal-based methods based on envelope variance (EV) and universal background model likelihood (UL). We also consider back-end approach based on N-best hypothesis (N-Best).

Envelop-variance measure for CS
The distortion measure proposed in [7] is extracted directly from the speech signal and it is based on the idea that reverberation smooths the time sequence of speech energy values, so the effect of reverberation may be observed as a reduction in the dynamic range of that envelope. In practice, a vector consisting of the estimated variances of compressed filter-bank energies (FBE) is obtained for each channel in each sub-band. The weighted average variance over all sub-bands is computed as an indicator of the amount of reverberation in each channel: the larger the envelope-variance measure is, the less reverberated is the respective channel.

UBM likelihood measure for CS
A new measure for CS that can be partially considered signal- and model-based has been explored. The proposed selection method is based on acoustic likelihoods, but rather than obtaining them from the ASR decoder for the particular recognized hypothesis, a general speech model is used to compute these likelihoods. The concept is similar to the universal background model (UBM) that is common practice in speaker verification [13]. In short, a Gaussian Mixture Model (GMM) is trained with all the data used for training the acoustic models, which is expected to model the general characteristics of the training corpora. Then, in the test phase, the likelihood obtained by each channel with this UBM model is computed. The working hypothesis is that the channels that obtain higher UBM likelihoods will match better the acoustic models and they will likely obtain better recognition performance. In this work, a GMM of 16 mixtures was considered, given that experiences with more mixtures showed no significant improvements.

N-Best hypothesis measure for CS
The likelihood normalization factor proposed in [10] is based in confidence measuring. For each channel the recognized probability of the $n^{th}$ best hypotheses is obtained. The confidence measuring is obtained by the ratio between the probability of the best sentence and all others hypotheses. Finally, the best channel corresponds to the highest measure of confidence.
4.3.3. Experimental results

Table 3 shows results exploiting the automatic channel, cluster and room selection strategies previously described. In the first and fourth rows, the results correspond to a random selection of the microphone inside the room that has been automatically identified following the majority voting strategy. Comparing to the results in the second row of Table 1, we observe a constant performance drop due to misidentification of the room. Anyway, the results are still much better than using a single reference microphone for all the events in the house. The second and fifth rows in Table 3 show the ASR performance when the microphone is automatically selected for recognition based on the highest EV or UL measure, respectively. The analysis of these results leads us into believing that these methods are not only able to provide meaningful room identification, but also to select individual well-performing microphones. Particularly, the results obtained with the EV measure are remarkably good. None of these approaches makes use of prior knowledge about the position of the sources or about the microphones location. The third and sixth rows show the performance achieved by the CS methods when they are used to select for each command a specific delay and sum beamformer, knowing the cluster and position of the speaker. While significant performance gains are obtained with respect to random cluster selection, the method does not outperform the simple delay and sum beamformer that processes all the microphones of each room.

The N-Best was only tested for the automatically selected microphone CS strategy. The N-Best WER is larger than the EV and UL WERs. This comparison shows that this method is not suitable for noisy environments and with multi-rooms.

## 5. MULTI-ROOM SPEECH/NON-SPEECH SEGMENTATION

Speech/non-speech segmentation of the acoustic input constitutes a crucial component of the DIRHA front end, providing important information to other system components, such as speaker localization, automatic speech recognition, and speaker recognition. For example, the speech recognizer does not need to operate on non-speech segments, thus improving robustness to noise and reducing computations. The module detects speech activity in the DIRHA environment and, with appropriate temporal smoothing and the possible use of multiple microphones, segments the audio stream into time intervals, classifying them as containing speech or not. In this paper, the system used for speech/non-speech segmentation was the MLP classifier using PLP features.

### 5.1. MLP Classifier Using PLP Features

This approach performs acoustic parametrization of the audio signal, extracting 12th order perceptual linear prediction (PLP) coefficients plus signal frame energy, all appended by their first temporal derivatives, thus yielding 26-dimensional acoustic features. These are subsequently passed to the classification block, which is implemented using an artificial neural network of the multi-layer perceptron (MLP) type [14]. The output of the trained neural classifier represents the probability of the audio signal containing speech. The last block of the speech/non-speech segmentation module is a finite state machine that receives as input the probability of the audio signal being speech. This block smooths the input signal, using a median filter over a small window. The smoothed signal is then thresholded and analyzed using a time window \((t_{\text{min}})\). The finite state machine consists of four possible states ("probable non-speech" non-speech, probable speech, and speech). If the input audio signal has a probability of speech above a given threshold, the finite state machine is placed into the probable speech" state. If, after a given time interval \((t_{\text{min}})\), the average speech probability is above a given confidence value, the machine changes to the speech state. Otherwise, it transitions to the non-speech state. The finite state machine generates segment boundaries for non-speech segments larger than the resolution of the median window. The value of \((t_{\text{min}})\) has been optimized to maximize non-speech detection.

### 5.2. Segmentation approaches for the whole house

The aim of this task is to detect the speech segments that occur in a room and, at the same time ignore the speech events that occur in the other rooms. In terms of evaluation, each speech segment detected in a room is considered to be correct if it has occurred in that room and regions of time with speech segments occurring in other rooms are ignored. In other words, the insertion errors due to detection of speech events belonging to events produced in other rooms are ignored in the evaluation of this task.

#### 5.2.1. Adaptation of MLP

The MLP model, described previously, does not work correctly with the acoustic properties of the ITEA apartment. A reasonable solution for this problem is of course to retrain or adapt the MLP based classifier using appropriate data, that is, data more similar to the test conditions. To evaluate the feasibility of this approach, the baseline MLP classifier was first adapted using three dev sets from DIRHASimCorpus, namely the ones in Italian (IT), European Portuguese (PT), and Greek (GR). As described in [5], the simulated data correspond to microphones located in five rooms of the ITEA apartment. For each room, a specific microphone was chosen (R1C, B2C, C1R, KA6, LA6). A total of 1125 audio files from the three languages, five rooms, and 75 recorded simulations were used in the

<table>
<thead>
<tr>
<th>CS measure</th>
<th>CS strategy</th>
<th>testing conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>dev</td>
</tr>
<tr>
<td>EV-based</td>
<td>room-auto random-mic</td>
<td>15.33</td>
</tr>
<tr>
<td></td>
<td>auto-mic</td>
<td>7.87</td>
</tr>
<tr>
<td></td>
<td>room-aware auto-beamf</td>
<td>5.62</td>
</tr>
<tr>
<td>UL-based</td>
<td>room-auto random-mic</td>
<td>16.49</td>
</tr>
<tr>
<td></td>
<td>auto-mic</td>
<td>15.05</td>
</tr>
<tr>
<td></td>
<td>room-aware auto-beamf</td>
<td>8.99</td>
</tr>
<tr>
<td>N-Best-based</td>
<td>auto-mic</td>
<td>27.42</td>
</tr>
</tbody>
</table>

Table 3. Average WER (%) performance exploiting envelope-variance (EV), novel UBM likelihood-based (UL) and N-Best hypothesis (N-Best) microphone selection techniques.
adaptation, of which 750 for training and the remaining 375 to validate the model. The MLP was fully adapted using a single epoch of backpropagation, with a much smaller learning step than the one used for the initial model training. This adaptation approach can be thought of as providing a best-case speech/non-speech segmentation, since the resulting classifier is being adapted on data collected in the exact same environment conditions (microphones and rooms) as the test set recordings.

5.2.2. Multi-Channel Approach

In this section, we exploit all the microphones available in the apartment to obtain an improved segmentation for each room. We explore two forms of multi-channel fusion: Majority Voting Decision Fusion (MVF) and Posterior Probability Fusion (PF). The MLP model adapted with the DIRHA-EP corpus is used in these two approaches.

Majority Voting Decision Fusion (MVF)

In the MVF all the microphones in the house are used and speech/non-speech segmentation is run individually for each channel. Then, the resulting segmentations from all the channels of a specific room are aligned to detect candidate speech events. Due to the different distances from the speech source to the microphones, the recordings have different time delays and consequently a time tolerance of 1 second is given to this alignment process. Then, if more than half of the microphones of a specific room detected a speech event candidate, it is considered that there was speech in that room.

Posterior Probability Fusion (PF)

In the PF the posterior probabilities obtained by the MLP for each channel of a specific room are combined before being processed by the state machine. The combination rule is simply the mean of the probabilities provided by the MLP. Then, the same state machine used in the single-channel case is used to obtain the room segmentation based on these averaged probabilities.

5.2.3. Evaluation Results For The Whole House

The results of the distinct approaches are presented in Table 4. In the mono-channel system, a representative microphone was chosen for each room (RIC, B2C, C1R, KA6 and LA6). The results are the mean of the detected segments in each room. It should be noted that the test data set of DIRHA-EP is not balanced, containing a much higher percentage of non-speech frames (75%) than speech ones (only 25%). Observing the speech recall values of Table 10, it can be seen that the MLP unadapted system rejects a very high percentage of speech (recall values are lower than 50%). After adaptation with in-domain data (MLP, DIRHA adapted system), speech recall increases to over 70%, while maintaining a high non-speech detection precision.

Despite the loss of some speech segments, a better performance was obtained for the techniques that used the set of microphones installed in the house. The non-speech segment detection results are quite similar for the different methods. This happens because the majority of the segments does not contain speech. Since the aim is to increase the precision of the system without the loss of speech segments, the best approach is the PF. The PF has higher precision and a drop in recall of only 0.8% of the segments when comparing with the MVF.

5.3. Room-Localized Segmentation

The aim of this section is the study of speech/non-speech segmentation that occurs uniquely in a specific room. In contrast to previous section, speech segments detected from other rooms are considered as insertion errors and affects the performance of the evaluated systems in this task. With this task, it is expected to assess the possibility of using the previous algorithms for simultaneous room localization and segmentation.

5.3.1. Results Without Room Selection

Baseline tests have been performed using the same previously applied approaches for this new task. That is, miss-detections due to speech events occurring in other rooms will affect the performance of the evaluated systems. The performance results in Table 5 are much lower than the ones in Table 4. This happens because many of the detected speech segments actually occurred in another room. Hence, one may conclude that the system is not effective for speech event localization.

5.3.2. Strategies for Room Selection

As stated previously, the segmentation algorithm does not perform a correct localization due to the errors introduced by the events in the other rooms. To fill this gap, we perform the speech detection with automatic room selection.

The proposed method is as follows. First, we obtain an automatic segmentation for each room using any of the previously described methods. With this operation we obtain speech candidate segments for each room. Then, speech candidate segments of all rooms are aligned with a tolerance of 1 second. This is done to match events that are likely to be the same ones, but that are simultaneously detected at different rooms. Finally, the last step is to decide to which room belongs every speech candidate segment using the information provided by an automatic detection room approach. From the various methods studied, the EV was chosen because it presents the best performance for an environment with noise and reverberation. In this case, the automatically detected room corresponds to the room to which belongs the microphone with the best EV measure in the region of the specific candidate segment. In practice, we have explored two methods of integrating the segmentation information and the room localization information. In a first approach, so-called EV (select_speech), we restrict the rooms in which the speech event may happen only to those rooms that actually detected that hypothesised segment. In a second approach, so-called EV (select_any), the automatic room detection is not restricted and the room localization can provide any room for each candidate segment. However, if the automatically selected room is not among the ones that actually detected the candidate speech segment, then that hypothesised segment is disregarded. In practice, the difference among the two methods is that in the first case, all aligned candidate segments are assigned to one room (and removed from any other that detected the same candidate), while in the second case, there may be candidate segments that are disregarded and not assigned to any room. Consequently, we may expect in the second approach an increase of the precision in exchange for a drop in the recall performance.

5.3.3. Results with Room Selection

Table 6 presents the results obtained for the two integrated approaches that combine speech/non-speech segmentation and room localization. Comparing these results with the ones obtained with
the baseline systems (Table 5), we can observe a great improvement in the precision performance of speech. On the other hand, there is also a significant drop in the recall performance. However, we can see that the introduction to room localization increases the system performance about 25% for the best method in terms of F-measure. Hence, it is demonstrated the convenience of the methods proposed combining segmentation and room localization.

With respect to the two approaches explored, in the first approach (select speech) the recall is higher, as expected, because all candidate segments are always assigned to one room. On the other hand, also as expected, the precision is very low when compared to the select any approach. In general, the second approach achieves a better generalised performance (F-meas).

### 6. CONCLUSIONS

In this work, we have presented a set of exploratory experiences addressed to analyse and to evaluate the performance of baseline speech processing components in European Portuguese for distant voice command recognition applications in domestic environments. For that purpose, a very realistic multi-channel and multi-room database named DIRHA-EP SimCorpus has been used. The analysis shows the importance of adequate room detection and channel selection strategies to obtain acceptable performances. In practice, according to our results, it seems that channel selection strategies can be more convenient than classical baseline beamforming methods –like the delay-and-sum– for this type of multi-room scenarios. We have explored two different computationally inexpensive channel selection measures and a decoder-based approach for room detection, channel selection and cluster selection. Experimental results show that the strategies based on envelope-variance measure consistently outperformed the remaining methods investigated. Finally, we have studied the problem of room-localized speech/non-speech segmentation, for which conventional segmentation approaches usually fail to provide consistent performances. We improved segmentation performance by combining multi-channel segmentation with envelope variance based room-localization. We explored two different approaches that permit obtaining different performances depending on whether it is more critical to lose or to miss-detect speech segments.

### 7. REFERENCES

<table>
<thead>
<tr>
<th>Type Approach EV</th>
<th>System Adaptation</th>
<th>speech</th>
<th>non-speech</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>select_speech</td>
<td>1c-MLP DIRHA-EP</td>
<td>65.9 43.2 52.2</td>
<td>92.9 97.1 95.0</td>
<td>90.9</td>
</tr>
<tr>
<td></td>
<td>MVF-MLP DIRHA-EP</td>
<td>65.3 46.4 54.3</td>
<td>93.8 97.1 95.4</td>
<td>91.7</td>
</tr>
<tr>
<td></td>
<td>PF-MLP DIRHA-EP</td>
<td>65.6 46.9 54.7</td>
<td>93.9 97.1 95.5</td>
<td>91.8</td>
</tr>
<tr>
<td>select_any</td>
<td>1c-MLP DIRHA-EP</td>
<td>59.5 73.2 65.7</td>
<td>98.2 96.7 97.5</td>
<td>95.3</td>
</tr>
<tr>
<td></td>
<td>MVF-MLP DIRHA-EP</td>
<td>59.6 75.2 66.5</td>
<td>98.4 96.7 97.6</td>
<td>95.5</td>
</tr>
<tr>
<td></td>
<td>PF-MLP DIRHA-EP</td>
<td>59.8 74.9 66.5</td>
<td>98.4 96.8 97.6</td>
<td>95.4</td>
</tr>
</tbody>
</table>

Table 6. Performance (%) of the “room-localized” speech/non-speech segmentation task applying single channel and multi-channel fusion approaches and combining with two different room-localization approaches based in EV.