An Overview of the ABCP1 System

`Technical Report TR-ABCP1-00`
`(DRAFT)`

Vitor M M C Pera

FEUP - Porto
June 2014

Abstract

This report presents an overview of the ABCP1 speech recognizer. A brief characterization of the speech applications that can be addressed by this system is initially presented. Then, the general structure and also some of the main features of the recognizer are discussed. More detailed information can be found in the referenced reports, which are going to be updated as needed, since for the time being the work to build the baseline recognizer is still in progress.

Contents

1 Introduction

The basic idea was to try and build an automatic speech recognizer that would be able to attain an WER bellow 5% on applications with the following main specifications: 1) continuous and paused reading speech 2) large vocabulary, up to 20K words 3) European Portuguese (EP) language 4) speaker dependency 5) use of acoustic and visual feature streams 6) real-time operation.

Well-known existing tools, such as the HTK[1] and the CMU/Cambridge-LM toolkit[2], allow to build efficiently some modules of the system, such as the audio-visual models and the language model. Other parts, in particular the decoder module, are built using in-house developed code. Although it is an harder approach, by the other side allows the full control of the recognition process and its better optimization based on the recognition task characteristics. Besides, regarding to the HTK, although being able to deal with multiple streams this toolkit is not appropriate to jointly decode audio-visual feature streams, in large measure due to the variable misalignment between the acoustical and the visual events. Moreover, it is intended to base the recognizer on two different types of sub-word units, phonemes in relation to the acoustics, as usual, and *visyllables* (corresponding to the visual realisations of the syllables) in relation to the visual speech. The implementation of techniques to mitigate the effects of the referred misalignment, such as relaxing product-HMMs for instance, and to support simultaneously phones and syllables, would imply extensive modifications of the code, in the case of the HTK free version. For the best of the author's knowledge, similar difficulties occur with other systems. Another quite popular open-source system is Julius[3], the standard speech recognition engine for R&D in Japan, which is based on a 2-pass decoder. This decoding process is suitable to the application here designed, mainly due to the large vocabulary and the different nature of the streams. By the other side, being a quite general purpose speech recognition platform, this system becomes uneasy to full optimization given a particular application, though it is highly configurable. As much as the author could foresee, this is specially true in relation to the audio-visual speech recognition (AVSR) approach, with some of the consequences already pointed out. Indeed, it seemed quite hard to change the code in order to integrate efficiently some of the existing knowledge on AVSR, both for improving the audio-visual modelling and to speed up the computations.

Briefly, the acoustic, visual and language models in the ABCP1 system, and also other components such as part of the visual analysis module, are built based on existing tools; and most of the developed code is dedicated

[1] http://htk.eng.cam.ac.uk

[2] http://mi.eng.cam.ac.uk

[3] http://julius.sourceforge.jp/en_index.php

to the decoder and to the acoustical front-end building, besides many small programs needed in other parts of the system. Part of all these pieces of code were easily adapted from an existing recognizer developed for the Resource Management (RM) recognition task[4].

To develop the ABCP1 system it was also created an appropriate data set, the ABCP-db1 speech database[1], which contains the needed acoustic and visual materials, besides other useful information. The raw materials in that data set were produced by one single subject, according to the application specifications, greatly simplifying the acquisition and processing tasks.

The main body of this report is divided as follows. Section 2 characterizes the speech recognition applications that can be implemented using the ABCP1 system. The structure of this system is presented in Section 3. The modules of this structure are briefly explained in Section 4. More details can be found in the referenced reports. It must be emphasized that the references concerning the techniques here alluded can be found in the References section in the respective technical report[5]. The conclusions are drawn in section 5. Given this is a technical report of a work that has not finished yet, it is included an Appendix with the maintained Log-book.

2 The speech recognition application

This speech recognition application, that was designed having in mind using it to support the development of the ABCP1 recognizer, has the following main properties:

- continuous speech, based on the European Portuguese (EP) language, spoken in a paused reading like manner;
- large vocabulary, up to 20K words;
- perplexity of the recognition task not atypical (approx. 100);
- speaker dependency, allowing users suffering of not severe disarthria;
- recognition based on acoustic and visual features extracted from speech;
- real-time operation, in a standard up-to-date (2010) PC;
- performance level corresponding to an WER bellow 5%.

[4] That recognizer was based on the simultaneous decoding of two different acoustical streams; even using a very simple language model, leading to a 2 digits perplexity, and acoustic models without context, its major drawback, 6% WER was achieved on a standard speaker independent, 1K word, test set.

[5] This means that all the used references on a certain "decoding acceleration technique", for instance, can be found in the References section of the tech. report "The ABCP1 decoder module".

Naturally, these specifications establish a reference for the kind of applications that the ABCP1 recognizer can address. Just a few comments about the items presented above must be made here [6].

A read-like continuous speech application was established mainly because this approach is natural enough for the intended purposes and, by the other side, it is obviously much simpler than recognizing spontaneous speech. One of the main reasons for the decision, from the very beginning, of using the EP language is based on the strong motivation of the author to study and to do some research on the audio-visual joint modelling of speech, in this language in particular (in spite of the much harder difficulties to get the appropriate data for the experiments). Just as an example, among quite a few other topics: very limited experimental work has been published in the audio-visual speech recognition (AVSR) based on the Portuguese language, and always were used words or visemes as linguistic units; besides, there are some results suggesting that an interesting approach, much beyond the direct improvement of the recognizers performance, would be based on visyllables[7] as the basic unit.

Concerning to the dimension of the vocabulary, it was set 20K words as the maximum value, so nowadays this application can be classified as medium to large in terms of the vocabulary. A smaller vocabulary would not be enough since it is intended that the ABCP1 recognizer can be used on applications such as text editing covering quite broad topics, enabling out of vocabulary words. On the contrary, increasing the vocabulary dimension much beyond 20K would naturally need a more complex decoder and certainly would increase substantially the WER if trying to preserve real-time operation.

Two factors were dominant on the decision of designing a speaker dependent application. Of course with less speech variability more easily can be developed the acoustic and visual models, demanding less data for training, and better performances can be expected even for somewhat smaller models. Another essential aspect consists on the strong motivation to use this application as an experimental platform to research and develop speech interfaces in the Assistive Technology area.

The use of visual features, besides the acoustical ones, was motivated by the research interests already noticed and also by their potential to improve the recognition performance. For some elementary sounds, and a classical example is the pair of phonemes /m/ and /n/, the visual cues bring relevant information that complements the acoustical cues, improving strongly its discrimination. By the other side, obviously the use of visual cues makes the recognition process more robust to disturbances that can occur in the

[6] More detailed information is available in the ABCP-db1 Dataset[1] and in the Language Model[2] technical reports.

[7] This could be the term meaning the visual correspondent to the syllables.

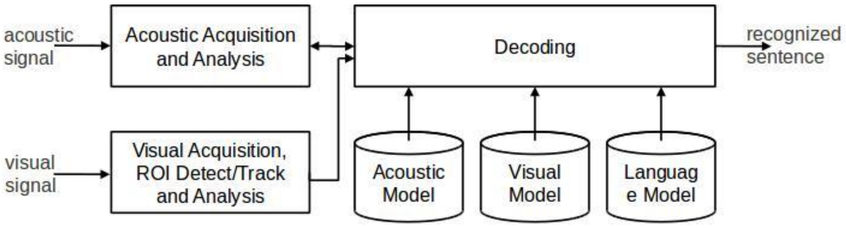

Figure 1: Block diagram of the ABCP1 system.

acoustic channel, generally a major problem in the ASR technology.

3 The structure of the ABCP1 system

Figure 1 presents the main structure of the system.

The acoustic acquisition module[6] allows the recognizer to operate both in the offline or in the on-line modes. The acoustic analysis is based on a standard cepstral representation of the signal.

For the time being, only offline operation mode is permitted by the visual acquisition and ROI detection/tracking module[7]. The captured image of the lips region is processed by the visual analysis module, which can be configured to combine a standard pixel-based technique with a very simple geometric approach.

The phoneme is the linguistic unit at the basis of the acoustic model (AM)[4]. Approximately 1K tri-phones are used in the AM, besides the silence and the garbage models. A much smaller set, with 38 mono-phones, is also used by a rough AM needed to accelerate the recognition process.

The visual model (VM)[5] is based on vi-syllabic units[8], numbering in the order of the hundreds.

The language model (LM)[2] uses two single pronunciation lexicons, covering the whole vocabulary: a standard phonetically based lexicon, and one lexicon based on syllabic units. At the syntactic level, two approaches are followed: in the initial decoding pass, only a simple 1-Gram and partial information from the word-pair restrictions can be applied; in the final pass, higher order nGram probabilities, with n equal to 2 or 3, are used.

The decoder[3] is based on a two-pass approach. In the first pass no visual cues are used and the LM is quite simple, such as already alluded, allowing a relatively easy implementation. The search is based on the Viterbi algorithm. The second pass is ignited according to the speech acoustics and is also conditioned by certain temporal restrictions. The stack decoder algorithm is then used to find the best recognition hypothesis, among those

[8] The visual realization corresponding to the syllabic units.

that survived to the initial pass. Both the acoustic and the visual feature streams are considered, besides the complete LM.

4 The main modules of the ABCP1 system

Each one of the following sub-sections emphasizes the main aspects of the ABCP1 modules, according to the blocks diagram presented in the Figure 1. More detailed information can be found in the respective technical reports (see the Reference section).

4.1 The speech acoustics acquisition and analysis

In the case of the on-line mode, the system uses the RtAudio[9] platform for real-time audio input. By default, a mono channel is open with 16 KHz sampling rate and 16 bit per sample. The signal is pre-emphasized and framed using a Hamming window with typical length and overlapping values. Every 10 ms, approximately, Mel-frequency cepstral coefficients (MFCC) and a non-normalized log-energy term is extracted from the signal. The first derivatives of the MFCCs and energy become also part of the features vector.

The computed feature vectors are not immediately submitted to the decoder. They pass through a FIFO buffer with size between 10 and 50 elements. This buffer acts like a previewing window, before decoding, on the acoustic signal, so that the ends of the decoding passes can be better decided. The recognizer can be configured so that information produced in the decoder is also considered in this decision. It is important to notice that the introduction of this delay, shorter than one half-second, in the recognition process, is perfectly acceptable given the nature of the application. Otherwise, a less simple optimized solution should be necessary.

In the case of the offline version of the acquisition and analysis module, the samples must be available in a WAV file. The log-energy term can be normalized and the cepstral mean normalization technique can be applied too. By default, the temporal restrictions used in the buffering mechanism are relaxed, so that pruning techniques can be soften or completely disabled.

4.2 The speech video acquisition and analysis

The existing implementation of this module allows to extract and save visual features from an AVI file. Those feature vectors can then be processed, jointly with the acoustic stream, when running the second-pass of the decoder. At the moment, does not exists a real-time version of this module,

[9] `http://music.mcgill.ca/~gary/rtaudio/`

mainly because the procedures and the code were not optimized in terms of speed[10]. The OpenCV 1.0.0 library[11] is at the basis of the existing code.

The frames in the video file contain the face of the applications user, who must stand relatively stable in front of the video camera. Besides, the video must be quite clean of disturbing factors such as variable shadows or sensible camera shaking. These restrictions are due to the fact that the existing baseline procedures for detecting and tracking the region of interest (ROI) are not robust enough, yet.

The frame rate is near 20 fps and also typical values in this kind of application[7] are used for other parameters of the image. The ROI, that is, the lips region, is initially detected based on the "Haar classifier cascade" approach. This classifier can be applied directly to detect the ROI, or else, can be trained to detect the nose and the eyes, which seem to be easier objects to locate, and then the lips can be more easily detected.

Given the initially located ROI and the imposed restrictions to the pose of the user, the tracking of the ROI is not a difficult task. It must be referred that the existing tracking module still is quite inefficient since does not fully considers those restrictions.

The typical size of the ROI is around 32x32 *pixel*, so strong compression is needed to get an affordable features vector. Actually, the analysis presents two kind of features. The pixel based approach uses the 2-dimensional discrete cosine transform (DCT2). In the default configuration, 24 low order coefficients, selected to allow better resolution along the vertical axis, compose each features vector. Reported results on audio visual speech recognition (AVSR), besides experiments on the reconstruction of the image from these coefficients, suggest this size is enough to carry most of the discriminative information in the video signal. It must be stressed that the application is speaker dependent, so greatly facilitating the analysis task. The other approach is based on geometric parameters extracted from the ROI. The existing version is very simple, just considering two measures: the horizontal and the vertical sizes of the ROI rectangle. The pixel based or the geometric features can be used isolated in the visual stream, or else, both can be combined to compose the features vector.

Since the visual stream is not used until the second-pass of the decoding process begins, during some speech segments the visual analysis could be properly conditioned by the linguistic classes on the local search space, according to the acoustic features[12].

[10] It is intended to develop such a module, given the importance of having an on-line version of the ABCP1 system.

[11] `http://www.vision.cis.udel.edu/opencv/`

[12] Preliminary experiments done suggest an interesting potential of this idea.

Symbol	Example	Art. manner	Art. place	Lips	Stress
a	pá	oral	back/low	-	yes
6	cama	oral	center/mid	-	-
6∼	mão	oral+nasal	center/mid	-	-
E	pé	oral	front/low	-	yes
y	bem	nasal	front/low	-	yes
e	dedo	oral	front/mid	-	-
e∼	tempo	oral+nasal	front/mid	-	-
@	dedal	oral	front/high	-	no
i	ida	oral	front/high	-	no
i∼	indo	oral+nasal	front/high	-	no
O	pó	oral	back/low	round	yes
o	bolo	oral	back/mid	round	-
o∼	som	oral+nasal	back/mid	round	-
u	tudo	oral	back/high	round	-
u∼	mundo	oral+nasal	back/high	round	-

Table 1: Vowels at the basis of the AM.

Symbol	Example	Artic. manner
j	pai	oral
w	mau	oral
w∼	mão	nasal

Table 2: Glides at the basis of the AM.

4.3 The acoustic model

The acoustic model (AM) is phonetically based and uses the hidden Markov model (HMM) as the essential statistical formalism.

In the actual version of the ABCP1 system, 38 phonemes constitute the elementary linguistic units at the basis of the AM. The considered vowels (#15), glides (#3) and consonants (#20) are shown in the Tables 1, 2 and 3, respectively. These Tables include information on several articulatory traces that can be useful when treating the visual modeling too.

Besides these sounds, two other basic classes are used in the AM, corresponding to the silence, or else to the noisy speech segments.

The AM of the ABCP1 recognizer is based on two sets of elementary models, which are built using HMMs. The main set contains approximately 1,000 HMMs, almost all modeling with context any of the 38 phonetic units already presented. For each phone, both left and right most important contexts are considered, establishing the set of tri-phones. The main set

Symbol	Example	Art. manner	Art. place	Voicing
b		oral stop	bilabial	voiced
p		oral stop	bilabial	voiceless
d		oral stop	alveolar	voiced
t		oral stop	alveolar	voiceless
g		oral stop	velar	voiced
k		oral stop	velar	voiceless
m		nasal stop	bilabial	voiced
n		nasal stop	alveolar	voiced
J	vinho	nasal stop	palatal	voiced
v		fricative	labiodental	voiced
f		fricative	labiodental	voiceless
z		fricative	apical	voiced
s		fricative	apical	voiceless
Z	janela	africate	palatal	voiced
S	chama	africate	palatal	voiceless
L	ilha	lateral	palatal	
l	lote	lateral	alveolar	voiced
h	igual	lateral	alveolar	
r	comer	vibrant	alveolar	
R	roda	vibrant	alveolar (multiple)	

Table 3: Consonants at the basis of the AM.

also contains di-phones and a few mono-phones also needed during the initial decoding pass, when no cross-word context is used. The other set is based on the HMMs that model, less accurately and without context information, the acoustic realization of the phonemes presented in the Tables 1-3. Moreover, in order to speed up the computations, first- and second-order derivatives are not used in the acoustic features. This second set is intended to allow the implementation of certain techniques[3] dedicated to the acceleration of the acoustic likelihoods computation. In both sets, besides the phonetic models also exist the silence model and the garbage model.

During the recognition, for each acoustic frame are initially computed the mono-phones emissions, following a state based approach. This computation can be accelerated using appropriate heuristics that also consider information from the previous frames. Then, the much more costly emissions of the tri-phones mixtures are selectively computed. Therefore, for great number of mixtures these emissions are simply estimated from the corresponding without-context models.

Each class is modeled through a semi-continuous HMM (SC-HMM), both in the mono- or in the tri-phone cases. Obviously, in each case quite different compromises between modeling accuracy and complexity must be established. For instance, concerning to the dimension of the codebook, even considering this is a speaker dependent AM, several thousands of gaussians can be expected in the case of the tri-phones SCHMMs and just a few hundreds in the case of the mono-phones.

Following a standard approach based on the progressive refinement of the models, those with context are built based on the mono-phone models. The HTK platform and other software tools support that development.

4.4 The visual model

The visual model (VM) is based on a set of elementary units associated to the visual realization of the syllables.

The syllable is better than smaller linguistic units such as the phoneme also because it is more robust to the context. This aspect can be even more sensible in the case of the visual features, when comparing to the acoustics. The cost of this advantage is, obviously, the quite larger number of elementary units. In the Portuguese language can be identified 3 or 4 thousand syllables. Fortunately, the number of visyllables is much lower since the visual realization of different syllables often leads to the same visyllable. The number of the different visyllables found in a ABCP-db1 data subset, containing approximately 10K words, is in the order of the hundreds (the selection of these units has not finished yet). Anyway, this number still is prohibitive when trying to model contextual information. So, the visyllables are modeled without context.

These visual patterns are established following the procedures: initially, the vocabulary words are syllabified based on some acoustic principles; and then, using a phoneme to viseme conversion table, a vi-syllabic vocabulary can be built. Besides, using available visual material and known techniques, the vi-syllabic vocabulary can be optimized considering the trade-off between size and discriminability.

Since the VM is used at the second decoding pass, when local restrictions in the search space due to the acoustic stream already are in effect, appropriate known techniques for using the visual stream can be followed. The use of these techniques imposes special care when building the VM. This area presents a great potential for trying innovative approaches[13].

Such as in the case of the AM, the SC-HMMs modeling these visyllables are developed based on the HTK platform.

4.5 The language model

The Language model (LM) has several components that are exclusive to each pass of the decoding process, in such a way as two LMs can be considered, the LM-1 and the LM-2, running in the first- and second-pass, respectively.

Considering also the lexical level, the LM-1 main components are: 1) a phonetically based single-pronunciation lexicon; 2) a unigram; 3) partial restrictions due to a word-pair grammar. The phonetic vocabulary consists of the 38 basic units presented in the Section 4.3. The lexicon covers a vocabulary size that can reach 20K words. The vocabulary is closed, according to the main recognition task defined in the database that supports the development of this recognizer. It is intended to implement an open vocabulary approach in a near future, following existing vocabulary optimization techniques[14] and filler models. The phonetic transcription of the words has an hybrid organization, combining a lexicon tree with a linear lexicon for the most common words in the vocabulary. The unigram probabilities are factorized along the lexicon branches. Given the way the lexicon is organized and also the process of establishing the candidate words, as the result of the initial decoding pass, the word-pair restrictions have just a partial effect on the search space restriction.

It is important to bear in mind that when the second-pass of the decoder is running the search space is based on a set of candidate words, or word trellis. The LM-2 main components are[15]: 1) a phonetic lexicon; 2) a visyllabic lexicon; 3) a 2-Gram or, optionally, a 3-Gram. A semi-automatic procedure is implemented leading to the definition of the visyllabic vocabulary, which size is in the order of the hundreds (see Section 4.4). Based on

[13] At this moment just a few ideas exist that seem appealing enough to be pursued.

[14] Using the same corpus, open-vocabulary recognition tasks are already established using these optimization techniques.

[15] Later, it is intended to extend the LM to restrictions based on morpho-syntactic and also semantic information.

these units, it is built a linear lexicon. At the syntactic level, the LM uses a 2-Gram or, optionally, a 3-Gram. Several configurations, including different discount techniques and cutoff parameters, can be established when building these grammars. Preliminary results using these grammars on the main recognition task in the ABCP-db1 dataset can be found in the technical report concerning the LM[2].

4.6 The decoder

Among the factors that lead to the decision of implementing a two-pass decoder, the following can be emphasized: 1) the existence of two feature streams of very distinct nature and, above all, consequently presenting quite different linguistic discriminability and temporal resolution (naturally, being the acoustic stream more discriminant and less coarse than the visual one) and, besides, with a natural and variable misalignment between the acoustic and the visual events, adapts well to this approach; 2) many published results on applications and recognizers presenting important similar aspects to the ABCP1, show the effectiveness of this solution to the information fusion problem; 3) the real-time operation capability and the notable recognition performance of the Julius recognition platform (see Section 1), based on a 2-pass decoder, in recognition tasks with similar important specifications, such as the vocabulary size and speaking style properties, allows to expect good results following this approach.

Nevertheless, it matters to say that since not all the available knowledge on the application is used in an unique pass and the visual stream is initially discarded, makes that the initial search space is larger than it could be if all the restrictions were simultaneously applied. By the other side, these options can simplify substantially decoding procedures, leading to an overall operation that can be relatively fast if appropriate acceleration techniques and pruning thresholds are applied.

If follows a brief description of one decoding cycle implicating the two passes.

Assuming the pass-1 is just beginning execution, an acoustic features vector is fetched from the Acoustic Analysis module. Then, a token based implementation of the Viterbi algorithm executes one more iteration on the search space supported by an hybrid lexical structure were is compiled a quite simple LM-1, such as referred in the Section 4.5. Since the real-time operation is a critical specification, several techniques, a few of them briefly referred in the Section 4.3, are implemented in order to accelerate the computations. The word candidates, including related information, ending at the actual instant are collected in a data structure that can be called a word-trellis. These steps repeat until the decoder receives, from the Acoustic Acquisition module, a signal to suspend the pass-1 and to start the pass-2. This signal depends on the speech acoustics and is also conditioned by certain temporal restrictions (the process can be configured so that this signal also depends on results obtained running the pass-1).

The pass-2 is based on the stack decoder algorithm, searching the best recognition hypothesis along the speech segment corresponding to the last execution of the pass-1. This pass uses the information in the word-trellis, including the AM and the partial LM scores, and also the visual stream, which scores can now be computed based on the VM. Besides, such as referred in the Section 4.5, at this stage are also used the complete LM restrictions. The search follows the inverse temporal direction, using the already existing AM scores and the LM partial scores in the searching heuristic function. According to published results, the acoustic scores can be partially recomputed allowing the relaxation on the words boundaries established in the word-trellis. In the condition that close instants can be associated to different recognition hypothesis, this should help mitigating severe consequences of (somewhat abusive) simplifications in the pass-1, in particular the searching bottleneck due to the lexicon tree lower nodes, which pernicious effects are reinforced by the simple one-best hypothesis assumption[16]. Given that the time available to execute the pass-2 can be short, at least in the worst scenarios (obviously, if very long speech pauses occur or if offline operation mode, there are no such problems), it is important that the VM scores can be fast computed, possibly imposing strict limits to the VM complexity and to the Visual Analysis procedures. When the execution of the pass-2 terminates, any new words sequence corresponding to the best recognition hypothesis must be made known.

5 Conclusions

In this report was presented an overview of the ABCP1 speech recognizer, including a brief characterization of the speech applications it can address, its general structure and also some of the main features. More detailed information can be found in the referenced reports.

[16] It sounds appealing experimenting this technique, though it substantially increases the complexity of the pass-2.

References

[1] V. Pera, ”The ABCP-db1 Data Set”, Tech. Rep., ABCP, Porto, (prev. 2015).

[2] V. Pera, ”The Language Model of the ABCP1 System”, Tech. Rep., ABCP, Porto, 2012.

[3] V. Pera, ”The Decoder of the ABCP1 System”, Tech. Rep., ABCP, Porto, (prev. 2015).

[4] V. Pera, ”The Acoustic Model of the ABCP1 System”, Tech. Rep., ABCP, Porto, (prev. 2015).

[5] V. Pera, ”The Visual Model of the ABCP1 System”, Tech. Rep., ABCP, Porto, (prev. 2015).

[6] V. Pera, ”The Speech Acoustics Acquisition and Analysis Modules of the ABCP1 System”, Tech. Rep., ABCP, Porto, (prev. 2014).

[7] V. Pera, ”The Speech Video Acquisition and Analysis Modules of the ABCP1 System”, Tech. Rep., ABCP, Porto, (2014).