

# Algerian Arabic Speech Database Project (ALGASD): Corpora's Elaboration

G. Droua-Hamdani \*, M. Boudraa \*\* and S. A. Selouani \*\*\*

\* Speech Processing Laboratory (TAP), CRSTDLA, Algiers, Algeria. Email: gh.droua@post.com.

\*\* Speech Communication Laboratory, USTHB, Algiers, Algeria. Email: mk.boudraa@yahoo.fr.

\*\*\* LARIHS Laboratory. University of Moncton, Canada Email: selouani@umcs.ca

**Abstract**— This paper presents different steps followed to realize Algerian Speech Database (ALGASD) for standard Arabic language.

The project concerns 300 Algerian native speakers whom are selected statistically from certain regions of the country and attempts to represent the principal variations of pronunciations denoted between the populations. Inspired by TIMIT protocol, ALGASD took into consideration the age and the education of each speaker. The basic text used to elaborate the database is constituted by 200 phonetically balanced sentences. Number of recorded sentences achieve to 1080 ones.

**Keywords**—Database; Algerian; speech

## I. INTRODUCTION

Considered as being the main resource of linguistic knowledge, the oral or written corpora are on the base of diverse remarkable works in various domains such: phonetics, phonology [1], sociolinguistics [2], automatic speech processing (Text-to-Speech, speech Recognition ...) [3], etc.

According to the aimed objectives and the language used to conceive them, several corpora are known and recognized for their big utilization. We quote as examples the most spread: TIMIT for the American English realized by the LDC for the evaluation of recognition systems [4], CHRISTINE spoken version of SUSANNE corpus and SEC, both for British English [5, 2]. Bref120 for French elaborated by ELRA [6], BDBRUIT realized in noisy environment for French [7], SpeechDat, multilingual project languages of the Eastern Europe concerning the domain of the telephony [8, 9], GlobalPhone, realized for 15 languages [10, 11], the multilingual base BABEL [12], CSTSC-Flight for the Chinese and which is reserved for the aeronautical domain [13], Indian languages: Tamil, Telug and Marathi [14], SPEECON containing 20 languages used in the development of voice-driven interfaces [15], POLYCOST dedicated to the applications of speaker recognition through the telephone lines [16], EUROM1 designed to aid the phonetic comparison of languages between seven European languages [17].

However the oral corpora interested by Arabic language in general (standard or dialects) remain less numerous comparing to the large number of Arabic speakers in the world (about 200 million). Indeed, we note only the corpora presented below: LDC corpus for the spontaneous telephone dialogues collected from Egyptian, Syrian, Palestinian, Lebanese and Jordanian speakers [18]. ELRA corpus for the Standard Arabic, read by Moroccans, Tunisians and Egyptians speakers [6]. The oral corpus of GlobalPhone realized from political and economic newspapers [10]. NEMLAR which is a speech corpus recorded from four different radio stations [19] and finally SAAVB for Saudi accented Arabic [1, 20].

## II. ALGASD REQUIREMENTS

ALGASD project consists on conception and realization of Algerian voice bank with the Standard Arabic as the substratum. Indeed, the speakers selected to record database are chosen among Algeria's regions.

### A. Notion of Standard Arabic

The Standard Arabic (SA) is compounded of 28 consonants and 6 vowels: 3 short vowels ([a], [u] and [i]) which are respectively ([fetʔa], [ʔamma] and [kasra]), with their 3 opposite long ones ([a:], [u:] and [i:]).

SA is characterized by many features used to distinguish between words. The most important ones are: the [madd] which corresponds to the long vowels, the emphasis and the germination. To occur, the emphasis needs a double articulation. This is necessary for the production of the emphatic phonemes which are: [©]/ص, [®]/ط, [‡]/ظ, [¼]/ض. Gemination means the reinforcement of a consonant's articulation which leads to the lengthening of phoneme's duration [21, 22].

### B. Algerian Languages

Situated in the north of Africa, Algeria extends over the vast territory of 2 380 000 km<sup>2</sup> occupied by about 34.8 million inhabitants whose majority are concentrated in the north of the country. The official language of the Algerians is SA, but their mother tongues are either Tamazight (Berber language) or specific variants of SA language which are stemming from the ethnic, geographical and colonial influences.

### C. Reference corpus

The reference corpus used for the ALGASD realization is based on [23] works which concern the elaboration of

200 Arabic Phonetically Balanced Sentences (APBS). The latter are organized into 20 lists of 10 sentences

In order to conceive them, the authors have used as foundations two important studies about Arabic roots and syllable distribution test done respectively by A.H. Moussa and P. Combescure [24 25].

### III. ALGASD TEXTS MATERIAL

From the reference corpus, we conceived three different sub-sets of texts material. Everyone aims to provide us a specific acoustic-phonetic knowledge. These new corpora constitute the basic start to elaborate our voice bank. Based on TIMIT protocol, the repartition adopted for each sub-corpus is as follow:

#### A. Common Corpus (Cc)

In order to list a maximum of dialectal variations of pronunciation observed among Algerians, the Cc corpus was read by all the speakers of the database.

This corpus has a characteristic to be composed of two sentences of APBS text which get firstly a largest number of Arabic phonemes and secondly a minimum of consonant's covering. Also, Cc presents the particularity to have the different Arabic consonants susceptible to present diverse regional pronunciations between Algerians such as [q] / [ق] and [¼] / [ح] sounds.

#### B. Reserved Corpus (Cr)

In order to bring all existing phonetic oppositions in the Arabic language such [d]/[t], [ʒ]/[ʒ], [ʃ]/[ʃ], [%]/[i], [ʻ]/[Y], etc. Cr corpus is endowed with 30 sentences chosen precisely from the reference corpus to provide us with this type knowledge.

However, to build Cr, we broke some times the organization and the balance of the text material in order to increase the frequency appearance of certain consonants. In fact, when we calculated the occurrence numbers of every phoneme of Cr, we noticed that the frequency appearance of some of them was very high in comparison with others. So, we replaced sentences which present redundant consonant by ones offering more infrequent outfit of phonemes.

#### C. Individual Corpus (Ci)

The last text material conceived for the voice bank is constituted of 168 remaining sentences from the reference corpus. By them, we wanted to obtain maximum of contextual allophones engendered by co-articulation phenomenon.

### IV. ALGASD CARACTERISTICS

To elaborate ALGASD, we selected 300 Algerian speakers from different regions of country. In order to get a representative and a realistic database, we have consulted the most recent census of inhabitants accessible in ONS web site [32]. According to the presented statistics, we choose for the study 11 representative regions which present the most important variation of pronunciations between Algerians

From the east to the west and from the north to the south, the regions taken into account are:

- Constantine, Jijel, Annaba from East;

- Oran, Tlemcen from West;
- Algiers, Tizi Ouzou, Médéa from Center;
- Bechar, El Oued, Ghardaïa from South.

According to population statistics, we distributed statistically 300 speakers of ALGASD on every region. All speakers are native ones and all living in their correspondent areas. The speaker profile used in database takes in consideration the age categories (18-30/ 30-45/ 45+) and the education levels for every speaker (Middle/Graduate/Post Graduate).

Table.1 shows exactly the distribution of speakers in each region according to sex of each one.

The recordings are made in quiet environments. The sound files have wave format, coded on 16 bits and sampling at 16 kHz

### V. ALGASD RECORDINGS STEPS

To record the voice bank ALGASD, we proceeded as follow:

#### A. Cr and Ci Recordings

We divided Cr into 10 sub-sets of 3 sentences which were distributed periodically on the 11 regions. After that, we added for each sub-set a unique sentence of C<sub>i</sub>. Finally, we achieved to 32 texts T<sub>i</sub> in a whole including 62 sentences onto 200 proposed in reference corpus (30 of Cr and 32 of Ci).

Moreover, to read T<sub>i</sub> texts, we designed for every region R<sub>n</sub> three speakers (2 male (M) and 1 female (F)) except for R9 where we endowed it only with 1 speaker of every sex because more than 2 speakers is statistically impossible. The total number of recorded sentences is equal to 128 sentences for 32 speakers.

We augment the sentences' number by enlarging the number of readings for every text T<sub>i</sub> by increasing the number of speakers for each region which becomes so 86 and sentences number to 344.

Speakers' distribution realized to record every corpus is explain in details in table .2. We can see from the latter that the repartition in the majority of region between female and male speakers approached 50% for every gender.

TABLE I. SPEAKERS' DISTRIBUTION IN SELECTED REGIONS

Regions	Rn	Nb. Loc. F	Nb. Loc. H	Total
Algiers	R1	40 (50%)	40 (50%)	80 (27%)
Tizi Ouzou	R2	17 (50%)	17 (50%)	34 (11%)
Médéa	R3	13 (52%)	12 (48%)	25 (8%)
Constantine	R4	13 (52%)	12 (48%)	25 (8%)
Jijel	R5	09 (50%)	09 (50%)	18 (6%)
Annaba	R6	09 (52%)	08 (48%)	17 (6%)
Oran	R7	19 (50%)	19 (50%)	38 (13%)
Tlemcen	R8	13 (50%)	13 (50%)	26 (9%)
Bechar	R9	04 (52%)	03 (48%)	07 (2%)
El Oued	R10	08 (50%)	08 (50%)	16 (5%)
Ghardaïa	R11	07 (50%)	07 (50%)	14 (5%)
<b>TOTAL %</b>	<b>11</b>	<b>152 (51%)</b>	<b>148 (49%)</b>	<b>300 (100%)</b>

TABLE II. SPEAKERS' DISTRIBUTION ACCORDING TO THE CORPUS AND THE REGIONS

Rn	Speak. .Nb. Cr		Speak. Nb. Ci	
	M	F	M	F
R1	12	11	30	29
R2	5	5	13	12
R3	4	3	9	9
R4	4	3	9	9
R5	3	2	7	6
R6	3	2	7	6
R7	6	5	15	13
R8	4	3	10	9
R9	1	1	3	4
R10	3	2	6	6
R11	2	2	5	5
<b>Total</b>	47	39	114	108
	86		222	

### B. Cc Recordings

The common sentences (Cc) are read by all speakers of ALGASD (300) which give to us a total of 600 recorded sentences.

Table.3 shows whole recorded sentences for every corpus in ALGASD according to the selected regions. We notice that the total number of recordings attains to 1080 sentences for the entire voice bank.

TABLE III. ALL RECORDED SENTENCES FOR EVERY CORPUS OF ALGASD

Rn	Cr	Cc	Ci	TOTAL
R1	69	160	59	288
R2	30	68	25	123
R3	21	50	18	89
R4	21	50	18	89
R5	15	36	13	64
R6	15	34	13	62
R7	33	76	28	137
R8	21	52	19	92
R9	6	14	7	27
R10	15	32	12	59
R11	12	28	10	50
<b>TOTAL</b>	<b>258</b>	<b>600</b>	<b>222</b>	<b>1080</b>

To sum up the repartition of texts material which mean Cr, Ci and Cc corpora, speakers' distribution and total recordings sentences according to corpus type; we propose the following table where all the results in details are explained.

TABLE IV. ALGASD ARCHITECTURE

Corpus	Sentences. Nb	Speakers. Nb	Record. Sentences
Cc	2	300	600
Cr	30	86	258
Ci	168	222	222
<b>TOTAL</b>	<b>200</b>	<b>300</b>	<b>1080</b>

## VI. CONCLUSION

This paper presents different steps followed to realize Algerian Speech Database (ALGASD) for standard Arabic language.

The architecture of the voice bank is inspired by TIMIT protocol.

Constituted by read sentences, the corpus is designed to provide speech data for the acquisition of acoustic-phonetic knowledge of 300 native speakers selected from 11 regions of Algeria. The selected areas are judged presenting the principal variations in the pronunciations of Algerian people. Our study is based on ONS statistics and 200 phonetically balanced sentences.

Total of recordings are 1080 sentences.

## REFERENCES

- [1] M. Alghamdi. KACST Arabic Phonetics Database. The Fifteenth International Congress of Phonetics Science. Barcelona, Spain. 2003.
- [2] L.J. Taylor and G. Knowles. Manual of Information to Accompany the SEC Corpus. Machine Readable Corpus of Spoken English. Available from: <http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM>
- [3] Edinburgh University Speech Timing Archive and Corpus of English (EUSTACE): <http://www.cstr.ed.ac.uk/projects/eustace>.
- [4] Texas Instrument and Massachusetts Institute of Technology corpus (TIMIT): Acoustic-Phonetic Continuous Speech Corpus. DMI. 1990.
- [5] The CHRISTINE project: <http://www.grsampson.net/RChristine.html>.
- [6] European Language Resources Association: <http://www.icp.grenet.fr/ELRA>.
- [7] J. Zeiliger & al. BDBRUIT une base de données "parole" de locuteurs soumis à du bruit. XX<sup>ème</sup> Journées d'Etude sur la Parole. Trégastel, 1-3 juin 1994.
- [8] H. van den Heuvel, V. Galounov, and H.S. Tropic. The SPEECHDAT (E) project: Creating speech databases for eastern European languages. Proceedings Workshop on Speech Database Development for Central and Eastern European Languages. Granada, Spain. 26 May 1998.
- [9] J. Cernocky and al. Recording of Czech and Slovak telephone databases within SpeechDat-E. Workshop on TEXT, SPEECH and DIALOG (TSD'99), Mariánské Lázně, Czech Republic, September 13-17, 1999.
- [10] T. Schultz and A. Waibel. GlobalPhone. Das Projekt GlobalPhone: Multilinguale Spracherkennung Computers, Linguistics, and Phonetics between Language and Speech, Bernhard Schröder et al (Ed.) Springer, Berlin 1998, ISBN Proceedings of the 4th

- Conference on NLP - Konvens-98, Bonn, Germany, October 1998.
- [11] T. Schultz and A. Waibel. Development of Multilingual Acoustic Models in the GlobalPhone Project Text, Speech, and Dialogue, Brno, Czech Republic, September 1998.
- [12] P. Roach, K. Vicsi. BABEL An Eastern European Multi-Language Database. COST249 meeting Zurich, 17-18 October 1996.
- [13] T.F. Zherng and al. Collection of a Chinese Spontaneous Telephone Speech Corpus and Proposal of Robust Rules for Robust Natural Language Parsing. Joint International Conference of SNLP- O-COCOSDA, Hua Hin, Thailand: 60-70, 2002.
- [14] A. Gopalakrishna & al. Development of Indian Language Speech Databases of Large Vocabulary Speech Recognition Systems. Proceedings of International Conference On Speech an Computer (SPECOM), Patras, Greece October 2005.
- [15] R. Siemund & al. 2000. SPEECON – Speech Data for Consumer Devices. Proceedings of LREC 2000.
- [16] D. Petrovska & al. POLYCOST: A Telephone-Speech Database for Speaker recognition. Proceedings RLA2C ("Speaker Recognition and its Commercial and Forensic Applications"), Avignon, France, April 20-23, pp. 211-214. 1998 (or: <http://circhp.epfl.ch/polycost>)
- [17] D. Chan and al. EUROM- a Spoken Language Resource for the EU. Eurospeech'95. Proceedings of the 4th European Conference on Speech Communication and Technology. Madrid, Spain, 18-21 September, 1995.
- [18] Linguistic Data Consortium: <http://www ldc.upenn.edu>.
- [19] K. Choukri, S. Hamid, N. Paulsson. Specifiacion of the Arabic Broadcast News Speech Corpus NEMLAR 2005: <http://www.nemlar.org>.
- [20] A.Mohamed, M. Alghamdi and Z. Muzaffar. Speaker Verification Based on Saudi Accented Arabic Database. International Symposium on Signal Processing and its Applications in conjunction with the International Conference on Information Sciences, Signal Processing and its Applications. Sharjah, United Arab Emirates. 12-15 February 2007.
- [21] J.F Bonnot. Etude expérimentale de certains aspects de la gémiation et de l'emphase en arabe. Travaux de l'Institut Phonétique de Strasbourg, N° 11, pp. 109-118, Strasbourg (France), 1979.
- [22] G. Droua-Hamdani. Durées des voyelles courtes -longues de l'arabe standard en milieux emphatique et gémigné, Colloque international en Traductologie et TAL, Oran (Algérie). 09- 11 Avril 2007
- [23] M. Boudraa, B. Boudraa et B. Guerin. Twenty Lists of Ten Arabic Sentences for Assessment. ACUSTICA Acta-acustica. Vol.84. 1998.
- [24] A.H. Moussa. Statistical study of Arabic language roots in Moejam Al-Sehah. Kuwait University. 1973.
- [25] P. Combescure. 20 listes de 10 phrases phonétiquement équilibrées. Revue d'acoustique 56 (1981) 34-38.
- [26] National Office of Statistics: <http://www.ONS.dz>.