OrienTel - Telephony databases across Northern Africa and the Middle East

Dorota Iskra[1], Rainer Siemund[2], Jamal Borno[2], Asuncion Moreno[4], Ossama Emam[3], Khalid Choukri[5], Oren Gedge[6], Herbert Tropf[7], Albino Nogueiras[4], Imed Zitouni[8], Anastasios Tsopanoglou[9], Nikos Fakotakis[10]

[1]SPEX, [2]ScanSoft, [3]IBM, [4]UPC, [5]ELDA, [6]NSC, [7]Siemens, [8]Lucent, [9]Knowledge, [10]University of Patras
c/o SPEX, Erasmusplein 1, 6525HT Nijmegen, the Netherlands
dorota@spex.nl

Abstract

OrienTel is a project that over the past two-and-half years developed speech databases and phonetic standards across Northern Africa, the Middle East and the Arabian Gulf. The project is funded by the European Commission and is coordinated by ScanSoft (Germany and Belgium). Other partners are ELDA (France), IBM (Germany), NSC (Israel), Siemens (Germany), Lucent (UK), Knowledge, the University of Patras (both Greece) and UPC (Spain), plus SPEX (the Netherlands) as validation agency. Now that OrienTel has passed the finish line, the present paper gives an update of the design conventions, and an account of the project's achievements. The paper also illustrates some of the challenges that the consortium faced which are mostly related to the validation and the research subjects.

Countries and technologies

The OrienTel area, ranging from the Gulf States in the East to Morocco in the West, is a very heterogeneous region. Its purchasing power in terms of GDP ranges from the UAE's US$ 16,800 to Yemen's US$ 304, technological infrastructure from Yemen's 1.4% fixed line teledensity to Cyprus' 54%. The number of mobile subscribers to mobile telephony networks stretches from virtually zero in Syria to close to 8 million in Egypt (CIT Publications, 2000). When OrienTel started in 2001, the consortium, consisting of ScanSoft (Belgium, Germany), ELDA (France), IBM (Germany, Egypt), NSC (Israel), Knowledge (Greece), Siemens (Germany), Lucent (UK) and the Universities of Patras (Greece) and UPC Barcelona (Spain), expected high growth rates particularly in the mobile part of the business. The database structure was therefore designed to meet the objectives of mobile speech applications via the telephone. By the end of the project, Arabic countries have surpassed the 30 million subscriber mark for mobile services, with an annual growth rate of 9.92%. For the first time the Arab countries have more subscribers to mobile services than to fixed lines. As 70% of all speakers were recorded via mobile networks, the OrienTel consortium is well equipped to meet the challenges the region poses.

Languages

From a linguistic point of view, too, the OrienTel region is very heterogenous. In order to treat the linguistic peculiarities of the area adequately, OrienTel followed a different strategy than previous projects of the SpeechDat type (Höge et al., 1999): each partner in the consortium was not responsible for a single language but for a whole country. The difference is an important one, since, as will be outlined below, in most OrienTel countries everyday-life is governed by more than a single language. One of the first project tasks was therefore to determine the various languages spoken in the OrienTel region, taking into account both linguistic and commercial criteria.

The varieties of Arabic spoken in the OrienTel region, to begin with, can be subdivided into four broad dialect regions, as outlined in Table 1. The consortium's idea was to cover each region in such a way as to be able to create both an acoustic model for each country and also a sufficiently representative Pan-Arabic acoustic model for the telephony speech market. The countries in italics are the ones actually covered by the consortium.

Dialect region	**Countries**
Maghreb Arabic	*Morocco*, Algeria, *Tunisia*, parts of Libya
Egyptian Arabic	*Egypt*, parts of Libya
Levantine Arabic	Syria, Lebanon, *Israel* + Palestine Authorities, *Jordan*
Gulf Arabic	Kuwait, Qatar, Bahrain, *UAE*, Saudi Arabia, Oman

Table 1: Dialect regions of Arabic

The linguistic situation in these countries, however, is a complex one. In Morocco, for example, the official language is Arabic. But this refers mainly to Modern Standard Arabic (MSA), the rather formal language of religion, the media and of public institutions. In everyday interaction, by contrast, people tend to speak a colloquial, essentially oral variant of Arabic that is only remotely related to the Standard (referred to herein as Modern Colloquial Arabic or MCA). Arabic is complemented by French as the language inherited from Morocco's colonial past and is heavily used especially in business interactions, in addition to the Berber, the native language of the North-Africa. All three (or even more) languages have their place in everyday life and user-friendly applications have to take into account each country's linguistic diversity and its users' preferences.

The databases produced in OrienTel and the number of speakers per database are depicted in Table 2.

Database specification

Due to the linguistic heterogeneity of the region, questions of database specification such as corpus composition, orthographic and phonetic transcription strategies constituted a crucial part of the project. Particularly Arabic and Hebrew posed interesting problems for speech recognition that were never tackled in projects of the OrienTel scale before. Cases in point are the rendering of

vowels, the right-to-left writing system and the transcription of the colloquial to standard continuum.

Country	1st language	2nd language	3rd language
Morocco	MSA 500	MCA 750	French 500
Tunisia	MSA 500	MCA 750	French 500
Egypt	MSA 500	MCA 750	English 500
Jordan	MSA 500	MCA 750	English 500
UAE	MSA 500	MCA 750	English 500
Israel	Hebrew 1,000	MCA 750	
Turkey	Turkish 1,700	German as spoken by Turks in Germany 300	
Cyprus	Greek 1,000	English 1,000	

Table 2: OrienTel language databases

Recording scenarios and platforms

All OrienTel databases were recorded from fixed and mobile networks via ISDN lines and multiple channels, i.e. either through a Basic Rate Interface or a Primary Rate Interface. A set of dialogues were implemented by the application driving the recordings, designed to make the caller speak and act comfortably in all the languages.

Corpus and vocabulary

Data collections relied on three separate sets of prompt sheets, namely one for Modern Standard Arabic, one for Modern Colloquial Arabic, and one for the 'business' languages (English and French in Arabic-speaking countries, plus Turkish, Greek, Hebrew and German).
While the specifications for English, French, Greek and German are largely based on the previous EC-funded SpeechDat and SpeeCon projects, the design for Arabic, Hebrew and Turkish presented a novelty. All three sets of prompt sheets, however, contained around 50 of the following items:

- isolated digits and number strings
- natural numbers and currency amounts
- yes/no questions
- dates and times
- application keywords and phrases
- word spotting phrases with embedded application words
- directory assistances names (proper names, place names, company names) and their spellings
- phonetically rich words and sentences
- spontaneous utterances

Transcription and annotation

The OrienTel transcription and annotation conventions were largely based on conventions used by the Linguistic Data Consortium and ARPA in producing the ATIS CD-ROMs, and the simplifications made for the SpeechDat-predecessors of this project, and SpeeCon. A coarse transcription was defined which could be performed quickly, but cover adequately the acoustic events most important for the training and testing of automatic speech recognisers. The transcription wass orthographic and included various markers representing audible acoustic events (speech and non-speech) present in the corresponding a-law files. All items for all languages covered were transcribed in standard orthography and romanised in the SAM label files (Gibbon et al., 1997).

Specification of speakers

As outlined in Table 2 above, the number of speakers recorded per country varied between 1,750 and 2000. Speakers were chosen according to certain predefined criteria.

Gender and age

The distribution of male and female speakers was 50% each per database, with an allowed deviation of 5%. There were no gender restriction for "Age" and "Dialect" (cf. below). For "Environment" (cf. also below), the gender distribution had to be 30-70% for each sub-category. Table 3 presents the distribution of speaker age.

Age	16-30	31-45	46+
Proportion	≥ 30%	≥ 20%	≥ 10%

Table 3: Distribution of speaker age

Dialect

Many (though not all) of the languages spoken in the OrienTel regions are not the speaker's actual mother tongue. In such cases, we consider persons who spent most of their childhood, or who grew up in the concerned region, as having no foreign accent. Language-specific cases were documented. Speech was collected from at least three different dialect regions - depending on the linguistic pecularities of each country, with at least 20 speakers recorded for each defined dialect. The speaker's dialectal region is determined by asking the question "*where did you spend most of your childhood*", not the question "*where do you live*". The allocation of city/district names to the corresponding dialect region was determined according to the information provided by each partner in the accompanying language-specific peculiarities document (the LSP document).

Distribution of environments

The speaker distribution for the mobile network was between 65 and 75% of the total number of speakers in the database. At least 30% of each gender had to be recorded in each environment. Both the fixed and mobile networks were further divided into specific environments. Speaker distribution in each environment is shown in Table 4.

	Environment	Speaker distribution
Fixed network 30% ± 5%	Home/office	≥ 75%
	Public place/Booth	(optional)
Mobile network 70% ± 5%	Home/office	≥ 20%
	Public place/Street	≥ 20%
	Vehicle	≥ 15%
	Hands-free car kit	≥ 5% (optional)

Table 4: Distribution of recording environments

Specification of the lexicon

The lexicon is an alphabetically ordered table of distinct lexical items that occur in the corpus with the corresponding pronunciation information. Each distinct

word has a separate entry, which is laid down in the order orthography (vowelised and non-vowelised for Arabic and Hebrew) ⇨ frequency ⇨ transliteration using SAMPA ⇨ phonetic transcription ⇨ pronunciation variants. The lexicon is derived from the annotated database.
The phonetic alphabet used is SAMPA. SAMPA inventories for the hitherto unavailable Arabic, Hebrew and Turkish sets were discussed and agreed on in close collaboration with Prof. Wells at UCL. They are available from http://www.phon.ucl.ac.uk/home/sampa.

Validation procedure

In order to guarantee an equally high quality of all the databases, a validation procedure was set up in a way similar to the previous SpeechDat projects (van den Heuvel, 1996). Except for internal quality checks by the partners themselves, an external institute (SPEX) was appointed to carry out the validation of all the databases. To provide the most efficient and useful feedback for the database producers, the validation was performed in a number of stages (van den Heuvel et al., 2004):

1. validation of the reading scripts (prompt sheets)
2. lexicon validation by an external expert
3. pre-validation of the first recorded 10 speakers
4. validation of the complete database
5. pre-release validation of master disks

In the various validation stages the following aspects of the database were checked either automatically or manually: documentation, formal structure and file names, corpus design, quality of speech files, the phonemic lexicon, orthographic transcription, speaker distribution, distribution of recording environments.
The above distribution of validation work in time was aimed at signalling problems at the earliest possible stage and giving the producers of the databases ample chance to compensate for the shortcomings before the recordings had reached an advanced stage.
The validation is carried out against a set of validation criteria which are derived from the specifications and which incorporate an extra tolerance margin (Iskra et al. 2002). The databases failing a number of these criteria have to be repaired before they can be accepted by the consortium.
In this way validation contributes to a high consistency and overall quality of speech databases produced by many different partners.

Validation challenges

The linguistic situation in the OrienTel countries posed challenges not only for specifications, but for validation as well. Just as there were three sets of specifications for different language variants, three parallel sets of validation criteria had to be designed in order to account for the differences in corpus content between the business languages, colloquial and standard variants of Arabic. The differences resulted, among others, from different prompting strategies. For business and colloquial variants, a high degree of freedom in the way of reading/uttering was given to the speakers by, for instance, presenting items containing telephone numbers as digits or eliciting spontaneous answers to questions. For standard Arabic, however, the speakers' tasks were more rigid by presenting all the items written out and restricting the degree of spontaneity to the minimum. For validation this implied varying degrees of control of the content of the items at transcription level.

Like in previous SpeechDat projects, the speech files were accompanied by SAM label files (Gibbon et al., 1997) containing various types of meta information as well as the prompt and the orthographic transcription of the utterance. In the orthographic transcription non-speech events such as background noise and speaker noise were marked as well as mispronunciations and truncations of the recordings. The fact that non-speech markers were indicated using Latin characters (from left-to-right) and the remaining speech transcription using Arabic alphabet (from right-to-left) made the word order in the mixed Arabic-Latin text unreliable. In order to circumvent this problem, an extra transliteration label was added to the label files containing SAMPA transliterations (only Latin characters) of the Arabic text. Transcription validation of non-speech markers, where the order is of vital importance, was, therefore, based on this transliteration label.

Another major problem posed by Arabic is the lack of vowel representation in written text that people are used to reading. As a result, the prompts were presented without vowels, which were only marked in exceptional cases in order to disambiguate. For the purpose of automatic speech recognition, however, vowels needed to be marked in the transcription. Moreover, the exact form of a given word was transcribed as it was pronounced by the speaker resulting in semi-phonetic transcriptions. These different strategies for the presentation of prompts and speech transcription led to severe discrepancies between the same words at the prompt and the transcription level. In order to relate the two without knowing the language (the knowledge which is not required for most parts of the validation) extra documentation effort was needed listing all the different pronunciation forms found in the database. For some checks this extra effort surmounted the value of the actual check and, therefore, for practical reasons was dropped.

The validation of the Orientel databases proved that the procedures which had been extensively used before had to be made still more robust in order to accommodate the peculiarities of Arabic languages.

Research

One of the goals of the OrienTel project was to show the feasibility of using the databases produced in the project in an ASR system and to improve their robustness across languages and dialects. For this purpose, some research tasks were developed.

Multilingual acoustic models and lexica

The objective was to explore and describe the potential of multilingual acoustic models and lexica. Rather than developing stand-alone acoustic models for one language, true multilinguality involves the phonetic rendering of more than one language variety in a single acoustic model and lexicon, accordingly. Two countries were chosen for

this experiment: Morocco with MSA, MCA and French, and Egypt with MSA, MCA and English.

Dialect adaptation

As a result of the collection throughout Arab countries, quite large databases have become available covering both MSA and MCA. These databases show a very good geographical, demographical and dialectal coverage and balancing. They cover extensively the linguistic status of six countries, where several languages and/or dialects coexist, and are widely used by a large part of the population. This makes these databases specially suited to investigate the possibilities of acoustic model adaptation between dialects.

Foreign accent adaptation

A third research task was the adaptation between native and foreign varieties of the same language, such as French in France and Morocco, English in Egypt and the UK, and German spoken by Turks and Germans. For this task, foreign accent adaptation of German spoken by Turks was addressed in this project.

With focus on Arab countries, the work concentrated on multilingual phonetic modeling for MSA, MCA and a foreign language, as well as the possibility of adaptation of an MCA variant, e.g., Levantine to another variety, e.g., Maghreb. Experiments were carried out under a common framework, i.e. training was done on phonetically balanced sentences and words, and test was done on digits, application words, and either times or dates. Each partner used their own in-house recognition system. It is well known that Arabic is usually written with consonants only and vowels are included while speaking. Vowelisation is not unique and depends, among others, on the context in which a given word is uttered. Table 5 shows an example. For each English term, as single application word was prompted, however, several correct realisations were recorded.

English Term	Prompt	Orthographic	SAMPA
<record>, <save>	سجل	سَجِّل	saZZil
		سَجَّل	saZZal
		سِجِّل	siZZil
		سجِّل	sZZil
		سَجّل	saZZl
<send>	صفط	صفِط	s`ift`
		صِفط	s`ifat`
		صفَّط	s`ffat`
		صفْط	s`fat`

Table 5: Example of multiple variations of the same word

As mentioned earlier a phonetic inventory was defined for each language and country. Research results proved the suitability of the defined set. For example, for Morocco a decision phonetic tree was automatically built. The results showed that, at the phone level, MCA and MSA are indistinguishable, the tree allowing phones from different languages to be together. However, context-dependent units showed a different behaviour. The context-dependent phonetic tree splits context units from different languages. This difference was evident in cross-language tests where word accuracy decreased significantly. ASR using multilingual MSA+MCA models improved slightly monolingual models both for MSA and MCA.

With respect to multilinguality, the multilingual models built in Egypt (Emam, 2004) with the three languages spoken there (MCA, MSA and English) and multilingual models made in Morocco (Moreno et al., 2004) with Arabic and French, showed very small degradation in the recognition scores, both for Arabic and the business language.
On the other hand, adaptation from one dialect (Levantine) to another dialect (Maghreb) by phone mapping and multilingual modelling (Saragosti et al., 2004) produced results very close to the recognition results from the monodialectal systems.

The results proved a great usability of all the resources generated in the OrienTel project to be adequately combined to build a general robust system for all the linguistic varieties spoken in the area.

Conclusions

The project finished officially in February 2004. Many databases are still under final validation, but all will be made publicly available through ELRA in due course.

References

AME Info (2004). Pan-Arabic mobile phone subscribers reach 30 million mark in 2003. Published online at http://www.ameinfo.com/news/Detailed/34111.html on January 27, 2004.

CIT Publications (2000). Telecommunications markets in the Middle East. Exeter: CIT Publications.

Emam, O. Multilingual acoustic models and lexica, focus on Egypt (2004). Orientel (IST-2001-28373) Deliverable D4.2.

Gibbon, D., R. Moore, R. Winski, eds. (1997). Handbook of standards and resources for spoken language systems. Mouton, de Gruyter. Berlin, New York.

Höge, H., C. Draxler, H. van den Heuvel, F.T. Johansen, E. Sanders, H. Tropf (1999). Speechdat multilingual speech databases for teleservices: Across the finish line. In Proceedings of EUROSPEECH '99, vol. 6 (pp. 2699--702). Budapest: ESCA

Iskra, D., H. van den Heuvel, O. Gedge, S. Shammass (2003) Specification of Validation Criteria. Orientel (IST-2001-28373) Deliverable 6.2.

Moreno, A, J.B. Mariño, A. Nogueiras, (2004). Multilingual acoustic models and lexica, focus on previous experiences and Morocco. Orientel (IST-2001-28373) Deliverable D4.1.

Saragosti, D., F. Gess, O. Gedge (2004). Dialect adaptation across the OrienTel region, focus on dialects of Arabic Orientel (IST-2001-28373) Deliverable D4.3.

Van den Heuvel, H. (1996). Validation criteria. SpeechDat Technical Report SD1.3.3.

Van den Heuvel, H., D. Iskra, E. Sanders, F. de Vriend, (2004) SLR Validation: Current Trends and Developments, Proceedings of LREC 2004.