

The Aix-MARSEC Project: An Evolutive Database of Spoken British English

Cyril Auran^{1,2}, Caroline Bouzon² & Daniel Hirst²

1 - IUFM d'Aix-Marseille / 2 - CNRS UMR 6057, Laboratoire Parole et Langage
Université de Provence, Aix-en-Provence, France
{auran; bouzon; hirst}@lpl.univ-aix.fr

Abstract

This paper presents the Aix-MARSEC project, an evolutive database of spoken British English. Specific details are given about the grapheme-phoneme conversion from the orthographic transcripts, the optimisation by elision rules of the phonetic transcription, the automatic phoneme-level alignment and automatic higher level treatment (syllables, subsyllabic structure, rhythmic units and MOMEL-INTSINT intonation coding). Integration of users' contributions will be within the general framework of GNU GPL licensing. Preliminary (pragmatic and prosodic) studies are presented in the final part of the paper.

1. Introduction

Freely available databases of spoken British English are scarce, not to mention such databases aligned with the speech signal at different levels ranging from phonemes to intonation units and intonation coding. The purpose of this paper therefore is to present the Aix-MARSEC project which aims to provide a solution to this availability issue. The project consists of two complementary parts: the Aix-MARSEC database (five and a half hours of natural sounding BBC recordings accompanied by multi-level annotation) and the Aix-MARSEC tools (a set of multi-platform Praat and Perl scripts and reference files, some of which are part of the PROZED project [10]). More specifically, this paper focuses on the processes and data format related to the Aix-MARSEC database. Emphasis is laid on the fact that one of the major characteristics in the project relies on its evolutive nature; indeed, all contributions from users considered compatible with the original technical and financial principles (GNU GPL) are to be integrated into the database.

Section two presents the origins of the corpus which constitutes the core of the database. Details on the phoneme-level treatments are then given in section three, which particularly focuses on grapheme-phoneme conversion, transcription optimisation by elision rules and automatic alignment. Section four deals with treatment at the levels of syllables, subsyllabic constituents, rhythmic units and intonation coding using the MOMEL and INTSINT automatic algorithms. Section five focuses on the actual format of the database. Section six, finally, gives examples of preliminary (pragmatic and prosodic) studies based on the database.

2. Aix-MARSEC: the origins

The SEC (*Spoken English Corpus*) is a collection of BBC recordings from the 1980s, grouping eleven different radio speech styles ranging from news and interviews to poetry reading ([15][16]). The data represents over five hours of natural-sounding British English from 53 different speakers (17 male and 36 female). The data includes about 55.000 orthographically transcribed words as well as a prosodic annotation (G. Knowles and B. Williams) using a series of fourteen tonetic stress marks.

The SEC was subsequently modified to facilitate computer use and became the MARSEC (*MACHine Readable Spoken*

English Corpus); the first change consisted in manually aligning the word and (minor-major) intonation unit boundaries with the sound. Second, some of the tonetic stress marks (TSM) were changed into ASCII symbols in order to have a computer compatible set of TSM ([19][20]).

We first converted the label text files into TextGrid format (Praat) and the word and intonation unit labels were then manually checked and modified in case of errors greater than 50ms.

3. Phoneme-level treatments

3.1. Grapheme-phoneme conversion

The 55.000 words of the corpus were automatically transcribed into phonemes from the orthographic transcription using Perl scripts: each word was looked up in an electronic dictionary, the *Advanced Learners' Dictionary* (with about 71.000 words), previously converted into the SAMPA alphabet.

The grapheme-phoneme conversion was also derived from a list of 700 manually transcribed words present in the corpus but absent from the pronunciation dictionary. A further stage consisted in treating the reduction of function words like *a*, *the*, etc. thanks to a list of words which were assumed to be reduced when not accompanied by tonetic stress marks.

The realisation of the morpheme '-s' (/s/ vs. /z/) of plural endings as well as that of the third person in the present tense were treated by means of a Perl function according to the phoneme context; the same sort of treatment was applied for the realisation of the morpheme '-ed' (/t/ vs. /d/) of the regular preterit and past participle forms. Numbers and dates received special treatment converting them into written words, followed by grapheme-phoneme conversion. Alphabetical abbreviations (of the type "BBC") were split into separate letters before being converted into phonemes. Mixed abbreviations and post codes (of the type "BBC2" and "Y01 1ET" respectively) received mixed treatment which separating numbers from letters, transcribing the former into written words, and converting the resulting word-letter sequence into phonemes.

In the phoneme-grapheme conversion, three dictionaries were consequently used: the general pronunciation dictionary (*Advanced Learners' Dictionary*), the 700 added forms and the list of reduced forms. All modifications or additions were made in separate dictionaries in order to ensure a possible subsequent update of the general dictionary without losing modifications. The output of this global phase in the grapheme-phoneme conversion consists in what we call the "raw transcription".

3.2. Conversion optimisation

The raw transcription was subsequently improved through the implementation of a series of twelve phoneme elision rules (cf. [3] for a fully detailed description), obtained from the observation of the corpus and from the literature [23], [13]

and [7]. The aim of the implementation of these rules is to improve the raw transcription but also the quality of the subsequent automatic alignment thanks to better adequacy between the signal and the transcription.

Following Campbell ([5]), the elasticity hypothesis is extended to the word level, a global z-score per word was then computed using Formula (1) below.

$$Word_Length = \sum_{i=1}^{Nb_phonemes} (mean_{phoi} + z^*sd_{phoi}) \quad (1)$$

A z less than a threshold τ (e.g. 0) can then be interpreted as a relative shortening of the word, constituting a sound basis for the implementation of the elision rules.

The elision rules are of two types:

- Three are “non phonotactic”: elision of phonemes whose predicted duration is below a threshold of 5ms and elision according to the morphological nature of words (for example, /h/ in the pronoun *he*);
- Nine phonotactic rules: elision of certain phonemes in specified phonemic contexts.

These rules were applied following three types of constraints: phonotactic constraints specified by each of the twelve rules, intonation constraints (rules were applied only in the absence of tonetic stress marks) and temporal constraints following the z-score of each word and a series of minimal phoneme duration thresholds taken from Klatt [14].

After the application of these rules, a total number of 4027 phonemes were eliminated from the raw transcription, leading to what we call the “elided transcription”.

The resulting phonetic transcription was evaluated by comparing it to a manual transcription using precision, recall and F-measure [21]. 74.4% of the elisions are well predicted (precision) and the application of the rules predicts 50.51% of the actual elided phonemes in the corpus (recall). Globally, the quality of the system concerning elision phenomena is evaluated to be around 60% (F-measure).

As a result of the optimisation method, the grapheme-phoneme conversion procedure yields a 94.79% agreement rate between the resulting transcription and the manual one. The recall measure however implies that approximately half of the elisions actually produced by the speakers are not currently predicted by the algorithm ([3]). New elision rules (some of which have been automatically generated during the evaluation phase) are thus to be introduced in the algorithm in order to improve the accuracy of the elided transcription.

3.3. Alignment procedure

The alignment of the elided transcription with the speech signal was carried out by Christophe Lévy and Pascal Nocéra from the Laboratoire d’Informatique d’Avignon using a classical Hidden Markov Model (HMM) and Viterbi algorithm forced alignment method ([22]).

As is common in automatic continuous speech recognition, non-contextual phonemes were modelled by a left-to-right HMM consisting of three emitting-states ([18]). For practical reasons of availability, each model was estimated on the TIMIT corpus, grouping 630 speakers (8 major dialects of American English) uttering 10 sentences. Following the TIMIT manual alignment, 48 phonemes were consequently learnt.

An HMM state was represented by a Gaussian Mixture Model with 8 components and diagonal covariance matrices estimated through the Expectation-Maximisation algorithm optimising the Maximum-Likelihood criterion. The speech signal was represented with 12 Mel Frequency Cepstral Coefficients (MFCC) obtained through filter bank analysis. The

12 MFCC were finally increased by the energy, the delta and delta-delta coefficients in order to obtain a 39-coefficient vector per speech frame.

3.4. Evaluation

In order to give a first approximation of the accuracy of the alignment method, the output of the forced alignment of the elided transcription was compared with the manual alignment of this same transcription. More particularly, phonemes given in the transcription but absent from the speech signal were reduced to 10 ms for better comparison with the automatic alignment (based on a 10 ms analysis frame). Table 1 sums up alignment accuracy as a function of acceptance-threshold [9].

Table 1: Accuracy of the alignment expressed as percentage of labels aligned at less than a given acceptance-threshold (AT) from the manual alignment

Acceptance Threshold	Elided transcription
64 ms	93.25 %
32 ms	82.02 %
20 ms	68.37 %
16 ms	59.97 %
15 ms	57.40 %
10 ms	42.43 %
5 ms	23.72 %

This evaluation, more particularly, shows that the alignment method used here yields an acceptable (though not excellent) accuracy score of nearly 70% for a 20ms AT.

Though this score already makes it possible to set up exhaustive experiments on the corpus, two solutions are to be tested in order to improve the quality of the system: indeed, resorting to British phonetic models and an alternative method of alignment relying on the iterative use of a Dynamic Time Warping procedure (based on the comparison of the re-synthesized speech signal with the original one [9]) may very well give better results than those observed so far.

4. Higher level treatment

4.1. Higher level

After being aligned with the signal, phonemes were automatically grouped according to the Maximal Onset Principle [17], which states that a maximum number of phonemes is grouped in onset position following onset phonotactic constraints specific to British English [7]. Following this syllabification, grouping and alignment into subsyllabic constituents (onset, nucleus and coda) were automatically calculated.

Syllables were subsequently grouped into rhythmic groups following Abercrombie’s [1] definition of the stress foot, as well as Jassem’s model of anacrusis and narrow rhythm unit [12]. According to these definitions, both foot and rhythm unit start with a stressed syllable; however, syllables in the corpus were not marked with lexical stress, a further stage thus consisted in automatically predicting such stressed syllables.

It was assumed that all words with more than one syllable had at least one stressed syllable; the stress patterns for these plurisyllabic words were automatically taken from the dictionaries along with the phonetic transcriptions. For the monosyllabic words, we used a list of monosyllabic function words which are by default always unstressed according to their position in the intonation unit (for example the pronoun *it* is normally unstressed in initial, medial and final positions). If not preceded by a TSM, these forms are considered unstressed and therefore considered as unstressed syllables in

the rhythmic grouping; if preceded by a TSM, they are by definition stressed and considered as such in the grouping. After the automatic assigning of the stresses, syllables are automatically grouped according to the two approaches.

Phonemes are also grouped into words according to their initial phonemic composition. All constituents are finally grouped into minor and major intonation units.

4.2. Intonation coding

The coding of intonation was carried out using the MOMEL-INTSINT methodology developed in Aix-en-Provence.

The MOMEL algorithm, more particularly, aims at modelling the actual F0 curve so that any microsegmental characteristics (the “micro-prosodic component”) should be factored out [8]. The resulting curve is thus similar to that found on a sequence of entirely sonorant segments and constitutes the “macro-prosodic component” [11]. The system uses a quadratic spline function which allow us to treat a sequence of target points as an appropriate phonetic representation of F0 curves.

The INTSINT algorithm automatically codes the sequence of MOMEL target points using a limited alphabet of abstract tonal symbols {M, T, B, H, L, S, U, D} standing for *Mid*, *Top*, *Bottom* (absolute tones), *Higher*, *Lower*, *Same*, *Upstepped* and *Downstepped* (relative tones) respectively. The INTSINT coding constitutes a surface phonological representation of intonation independent from any a priori phonological inventory of the intonation patterns of a given language.

Both these algorithms are called from within a Praat script and are available for the Microsoft Windows (98, 98SE, ME, 2000, XP), Intel based Linux and Mac OS X operating systems.

5. Data format

The Aix-MARSEC project has been designed as a freely available (GNU GPL) evolutive environment in which contributions from all referenced users could be integrated. As we mentioned earlier, the project relies on both the Aix-MARSEC database and the Aix-MARSEC tools, both aspects being very likely to benefit from users’ contributions.

The Aix-MARSEC database consists of two major components: the audio component (most of the original recordings from the SEC/MARSEC corpus) and the annotation component.

For compatibility and processing reasons, the 332-minute-long audio component is available under the form of 408 16 kHz .wav format files.

The annotation component currently comprises the 9 different levels mentioned earlier: phonemes, syllables, subsyllabic constituents, words, stress feet, rhythm units, minor and major intonation units, INTSINT coding and the corresponding values of the targets in Hz. Each level is represented by a separate tier in Praat TextGrids (as illustrated in figure 1). Two supplementary levels, based on the syntactic annotation of the corpus using the CLAWS system and a Property Grammar system developed in the Laboratoire Parole et Langage in Aix-en-Provence are to be integrated soon, thus allowing not only future analyses taking into account the grammatical tagging and parsing of the data, but also the direct comparison of automatic syntactic annotation systems.

The Aix-MARSEC tools consist of a set of reference files (grapheme-phoneme conversion dictionaries) and (multi-platform) Praat and Perl scripts. The scripts, fall into two categories: the first category (‘Aix-MARSEC tools’ proper) consists of tools specifically designed for the processing of the Aix-MARSEC database (selection of a subset of Praat TextGrid tiers, etc.); the second category consists in a subset of the ProZed project [10], applied to the Aix-MARSEC database

(signal processing, XML format conversion, etc.). All these tools are to be made freely available to the community for non-commercial non-military research.

New annotation levels, as well as new tools and references, contributed by referenced users are to be integrated into the Aix-MARSEC project provided they abide by its elementary technical (multi-platform open source software) and financial (free resource) principles.

6. Exploratory studies

6.1. Rhythm

Thanks to the different levels of annotation and alignment, a study is currently in progress concerning prosodic timing in British English [4]. Specifically, Abercrombie’s [1] and Jassem’s [12] models of phonological structure are compared (i.e. the alignment on the levels of stress feet and rhythm units/anacrusis) in order to determine which one better corresponds to the rhythmic structure of British English.

6.2. Anaphora and resetting

Annotation at the orthographic, syllabic and intonation (through MOMEL modelling in particular) levels has allowed automatic analyses in the study [2] of complex interactions between pragmatic and prosodic parameters regarding resetting phenomena in relation with anaphoric chains [6], discourse connectors and discourse topical structure [24].

7. Conclusion and perspectives

This paper presents the Aix-MARSEC project designed as a freely available evolutive database and set of tools (to be available from the following address.

<http://www.lpl.univ-aix.fr/~EPGA/>).

The database currently consists of approximately five and a half hours of recordings, annotated on nine different levels (phonemes, syllables, subsyllabic constituents, words, stress feet, rhythm units, minor and major intonation units, INTSINT coding and the corresponding values of the targets in Hertz).

The Aix-MARSEC tools comprise diverse multi-platform Praat and Perl scripts related either to the automatic transcription and annotation of the original data or to the conversion of the final data into more manageable sets (TextGrid tiers selection and/or ProZed project XML format conversion).

Further improvement in the overall quality of the database is envisaged: increase in recall rate (for the grapheme-phoneme conversion phase) using new elision rules and alignment optimisation through the use of adequate sets of phonetic models and iterative use of DTW.

In its present state however, the database already constitutes privileged material for diverse experiments and studies related to prosodic, linguistic and pragmatic aspects of spoken discourse.

Conceived as an evolutive project, Aix-MARSEC aims at integrating contributions from all users, thus leading to new levels of annotation and new tools for the automatic processing of natural sounding speech.

8. References

- [1] Abercrombie, D., 1967. *Elements of General Phonetics*. Edinburgh : Edinburgh University Press.
- [2] Auran, C ; Hirst, D.J., 2004, submitted. Anaphora, Connectors and Resetting: Prosodic and Pragmatic Parameters in the Marking of Discourse Structure. *Speech Prosody 2004*, March 23-26.

- [3] Auran, C. ; Bouzon, C., in press. Phonotactique prédictive et alignement automatique : application au corpus MARSEC et perspective. *TIPA*, 22, 13-44.
- [4] Bouzon, C.; Hirst, D.J., submitted. Isochrony and prosodic structure in British English. *Speech Prosody 2004*, March 23-26.
- [5] Campbell, N., 1992. *Multi-level Timing in Speech*. PhD Thesis, University of Sussex.
- [6] Corblin, F., 1995. *Les formes de Reprise dans le discours. Anaphore et chaînes de référence*. Presses Universitaires de Rennes.
- [7] Cruttenden, A., 1997. *Gimson's Pronunciation of English. Fifth edition*. England: Arnold.
- [8] Di Cristo, A. ; Hirst, D.J., 1986. Modelling French micromelody: analysis and synthesis. In Kohler (ed.): *Prosodic Cues for Segments (Phonetica 43, 1-3)*, 11-30.
- [9] Di Cristo, P. ; Hirst, D.J., 1997. Un procédé d'alignement automatique de transcriptions phonétiques sans apprentissage préalable. *4^o Congrès Français d'Acoustique*, 1, Marseilles, April 14-18, France : SFA, Teknea.
- [10] Hirst, D.J., 2000. ProZed: a multilingual prosody editor for speech synthesis. *IEE Workshop on Speech Synthesis*. London, April 2000.
- [11] Hirst, D.; Di Cristo, A.; Espesser, R., 2000. Levels of Representation and Levels of Analysis for the Description of Intonation Systems. In Horne, M. (ed.): *Prosody : Theory and Experiment. Text, Speech and Language Technology, 14*. Kluwer Academic Publishers, 51-87.
- [12] Jassem, W., 1952. Stress in modern English. *Biuletyn Polskiego Towarzystwa Językoznawczego IX*, 21-49.
- [13] Jones, D., 1990. *English Pronouncing Dictionary*. London: Longman.
- [14] Klatt, D.H., 1979. Synthesis by Rule of Segmental Durations in English Sentences. In Lindblom, B. and Ohman, S.E.G. (eds): *Frontiers of Speech Communication Research*, 287-299.
- [15] Knowles, G.; Wichmann, A.; Alderson, P., 1996. *Working with Speech: perspectives on research into the Lancaster/IBM Spoken English Corpus*. England: Longman.
- [16] Knowles, G.; Williams, B.; Taylor, L., 1996. *A Corpus of Formal British English Speech*. London: Longman.
- [17] Pulgram, E., 1970. *Syllable, Word, Nexus, Cursus*. The Hague : Mouton.
- [18] Rabiner L.R., 1984. A tutorial on hidden Markov Models and selected applications in speech recognition, *IEEE transactions on Speech Audio Processing*, vol. 2.
- [19] Roach, P.; Knowles, G.; Varadi, T.; Arnfield, S., 1993. MARSEC: A machine readable Spoken English corpus. *Journal of the International Phonetic Association 23(2)*, 47-53.
- [20] Roach, P., 1994. Conversion between prosodic transcription systems: "Standard British" and ToBI. *Speech Communication*, 15, 91-99.
- [21] Van Rijsbergen, C.J., 1979. *Information Retrieval, 2nd edition*. University of Glasgow.
- [22] Viterbi A., 1967. Error bounds for convolutional Codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, vol. 2.
- [23] Wells, J.C., 1990. *Pronunciation Dictionary*. London : Longman.
- [24] Wichman, A., 2000. *Intonation in Text and Discourse*. London : Longman

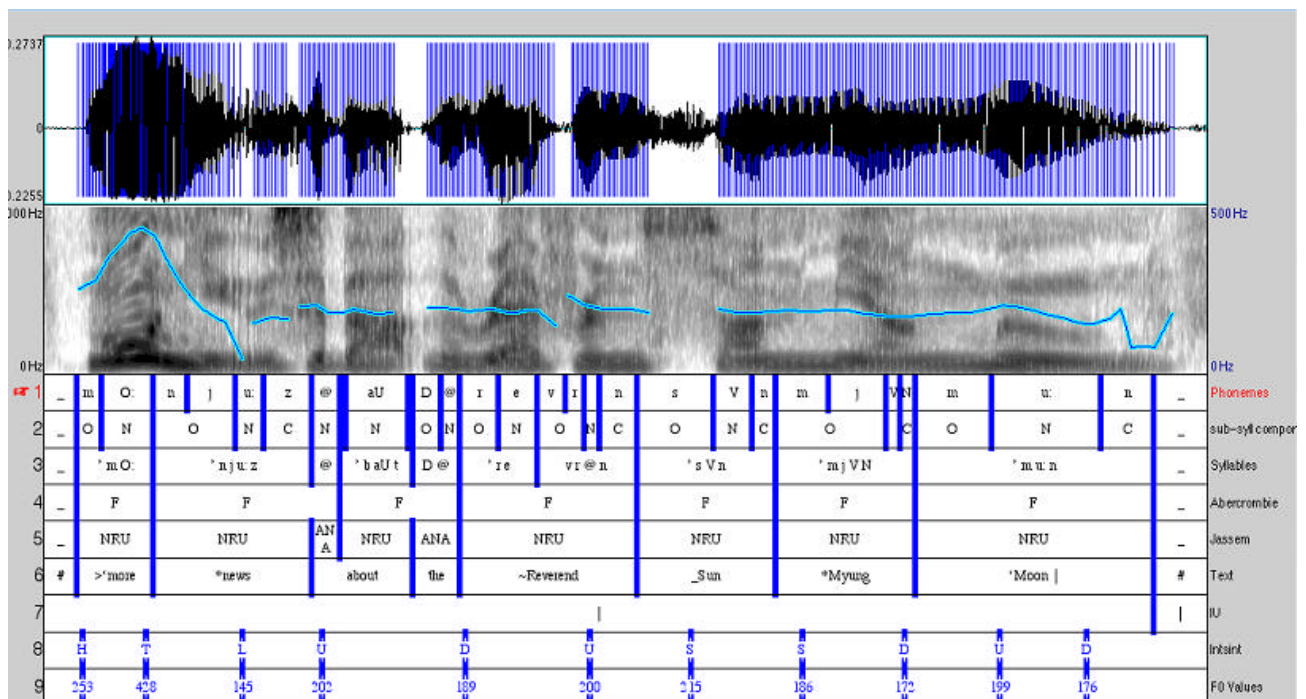


Figure 1 : illustration of the different levels of annotation of an extract from the corpus