Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5337–5342
Marseille, 11–16 May 2020

Arabic Speech Rhythm Corpus: Read and Spontaneous Speaking Styles

Omnia Ibrahim [1&2],Homa Asadi [1], Eman Kassem [2], Volker Dellwo [1]
[1] Institute of Computational Linguistics, University of Zurich, [2] Phonetics and linguistics department, Alexandria University
[1] Andreasstrasse 15, 8050 Zürich, Switzerland, [2] Qism Bab Sharqi, Alexandria, Egypt
ibrahim@ifi.uzh.ch, eman.qasem@alexu.edu.eg, homa.asadi@cl.uzh.ch, volker.dellwo@uzh.ch

Abstract

Databases for studying speech rhythm and tempo exist for numerous languages. The present corpus was built to allow comparisons between Arabic speech rhythm and other languages. 10 Egyptian speakers (gender-balanced) produced speech in two different speaking styles (read and spontaneous). The design of the reading task replicates the methodology used in the creation of BonnTempo corpus (BTC). During the spontaneous task, speakers talked freely for more than one minute about their daily life and/or their studies, then they described the directions to come to the university from a famous near location using a map as a visual stimulus. For corpus annotation, the database has been manually and automatically time-labeled, which makes it feasible to perform a quantitative analysis of the rhythm of Arabic in both Modern Standard Arabic (MSA) and Egyptian dialect variety. The database serves as a phonetic resource, which allows researchers to examine various aspects of Arabic supra-segmental features and it can be used for forensic phonetic research, for comparison of different speakers, analyzing variability in different speaking styles, and automatic speech and speaker recognition.

Keywords: Speech corpus, Arabic rhythm, Egyptian dialect, stress-timed language

1 Introduction

The successful collection of data is a key stage to obtain reliable and valid results in phonetic research. In this paper, we report on work-in-progress about the construction of a speech corpus for Arabic in Egyptian dialect. The primary intention behind designing such a corpus is to provide a homogeneous database of Arabic speech recordings, which investigates broadly/narrowly acoustic parameters of Arabic speech rhythm for forensic voice comparison (FVC) research and casework application. In a typical FVC casework, two samples of voices, a known and an unknown (disputed) sample, are compared to estimate the probability that the same speaker has produced the speech samples (same-speaker hypothesis) versus the probability that the speech samples have come from two different speakers (different-speaker hypothesis) (Rose, 2002). To objectively estimate this probability, having access to speech corpus containing samples from Arabic native speakers that contributes to the knowledge about between-speaker variability and within-speaker variability (Morrison et al., 2012; Kinoshita, 2001).

The motivation for exploring speaker-specific acoustic properties of speech of Arab speakers stems precisely from the lack of population statistics for the Arabic language in FVC. As yet, to the best of our knowledge, there is no forensically relevant Arabic speech corpus available being capable of utilizing in FVC research. The current corpus fitted for FVC research aims to fill this gap and it can be considered as the first database of its kind in Arabic which consists of high-quality audio recordings from a regionally and socially stratified population involving different speaking styles. The corpus follows principles of the protocol for the collection of databases of recordings for forensic-voice-comparison research and practice, which was developed by Morrison, Rose and Zhang (2012). Based on the protocol requirements, a dataset can be suitable for FVC research which fulfills three criteria: 1) non-contemporaneity of recording sessions for each speaker, 2) using different speaking styles for the recordings of each speaker, and 3) usability for research and casework involving recording and transmission-channel mismatch.

One of the promising and newly developed lines of investigation for FVC extracts speaker-specific information from the temporal organization of speech. Recent evidence from numerous different languages has shown that speech rhythm characteristics based on consonantal and vocalic durational variability as well as syllabic intensity have potential in capturing between-speaker variability (Dellwo et al., 2015; Leemann et al., 2014; He and Dellwo, 2016). The present study will, therefore, incorporate speech rhythm measures into the dataset by following the collection procedures of BonnTempo corpus (Dellwo et al, 2004), which is one of the databases currently available for the study of speech rhythm measures in connection with speech rate. The current corpus will thus be an extension to BonnTempo corpus, which subsequently allows us to investigate temporal characteristics of speech in MSA from both an acoustic and speaker-specific point of view in future projects.

1.1 Arabic language background

Arabic language (ISO 639-3: ara) belongs to the Semitic language family and it is the fourth most spoken language in the world with an estimated number of 400 million speakers over 23 countries as an official language (Bateson 2003). Hence, it has gained much attention from researchers in both phonetic description and development of speech synthesis and speech recognition fields.

There are a number of Arabic varieties; the first type, Classical Standard Arabic, which is the language of the Quran and classical literature. The second is Modern Standard Arabic (MSA). It is considered to be the modern version of Classical Arabic (Al-Sobh et al., 2015) and is the language of formal speech in Arab countries, such as is used in governmental speeches, the education system

and on the news. However, MSA is not the language used in everyday life and is considered a second language for all Arabic-speakers. Third, Colloquial (dialectal) Arabic is the language of everyday speech and conversation (Al-Suwaiyan, 2018). One of the colloquial Arabic is the Egyptian Arabic (the spoken variety of Arabic found in Egypt). Because of music and Media, most people in the Arab world understand Egyptian Arabic.

The phonological system of Arabic has 34 phonemes: 6 vowels (3 short vowels with 3 opposite long ones) and 28 consonants. Among these consonants, there are two distinctive classes, which are named pharyngeal and emphatic phonemes (Alghamdi 2003). The Arabic syllabic structure can be summarized in the following rule: CV(:)(C)(C) which mean there are three types of syllables in Arabic, light (CV), heavy (CVV and CVC) and super-heavy (CVVC, CVCC, CVVCC) (Watson 2011). For the suprasegmental aspect, Arabic is categorized as a stress-timed language. Furthermore, Word stress in Arabic is non-phonemic which implies that stress is not meaning-distinguishing. In MSA only the last three syllables of the word are relevant for determining stress, which means that stress never falls on the pre-antepenultimate syllable or before that.

1.2 Speech rhythm

Languages of the world are often classified into distinct rhythmic types of which the two most prominent are the stressed-timed and syllable-timed rhythm classes (Ramus et al., 1999; Grabe and Low, 2002). Stressed-timed languages are known to have higher durational variability of both consonantal and vocalic interval duration as well as a more complex syllable structure compared to syllable-timed languages. For example, English and German are considered to be a stress-timed language where they emphasize particularly stressed syllables at regular intervals, while French, on the other hand, appears to space syllables equally across an utterance (Ramus et al., 1999).

Acoustic correlates of speech rhythm are based upon different phonetic durational units (Ramus et al. 1999; Grabe and Low 2002) over syllables or feet (Nolan and Asu 2009), voiced and unvoiced intervals (Dellwo and Fourcin 2013) to amplitude peak intervals (Marcus 1981). Such rhythmic measures also belong to two domains pertinent to durational characteristics of speech and amplitude envelope. Acoustic measures of speech rhythm based on the durational characteristics of consonantal and vocalic intervals as well as the syllabic intensity can reveal between-language and between-speaker variability. These correlates have provided new insights into how speech timing functions both across and within languages and they have also been applied to developmental and pathological questions.

One of the speech corpora provided for the study of speech rhythm in different languages is the BonnTempo Corpus (BTC), which has been constructed by Dellwo et al. (2004). It consists of a read short story in 5 different speaking rates (normal, slowest, slow, fast and fastest) in 5 languages and 4 second language conditions, while the absolute number of speakers per language still varies considerably. The languages were selected to represent both traditional rhythmic classes. Stress-timing is represented by English and German, while syllable-timing by French and Italian. The corpus presented in this paper will be an extension to BonnTempo Corpus.

Practical rhythm studies related to Arabic are relatively less numerous compared to the studies dealing with other languages like English, Korean, French, Spanish and Portuguese Grabe (2003), Jang (2009), O'Rourk (2008). Several studies investigated the rhythmic pattern of different Arabic dialects; in their study (Ghazali et al. 2002), an acoustic investigation of the proportion of vocalic intervals and the standard deviation of consonantal intervals in six dialects (Morocco, Algeria, Tunisia, Egypt, Syria, and Jordan) was carried out. The subjects were 4 Moroccans, 2 Algerians and 2 Tunisian speakers representing Western Arabic, and 2 Jordanians, 3 Syrians and 1 Egyptian representing the Middle East. Their results show that complex syllable and reduced vowels in the Western dialects, and longer vowels in the Eastern dialects seem to be the main factors responsible for differences in rhythmic structures. In another study (Altuwaim et al. 2014), researchers used various timing metrics that have been suggested for quantifying rhythmic differences between Two Saudi dialects. Their dataset containing read sentences were created based on MSA rules. There are 62 audio files uttered by speakers from the Riyad region and 39 utterances in the Buraidah dialect. They investigated the use of rhythmic measures to discriminate between the Saudi dialects. Droua-Hamdani et al. (2010) investigated the Arabic rhythm of 73 Algerian speakers who read two sentences and they concluded that although Arabic is classified as a stressed-timed language, Algerian Arabic tends to be an intermediate language between stressed and timed languages. In their study, Hamdi et al. (2005) investigated the relationship between the syllabic structure of Arabic dialect and the rhythmic class they belong to. Their analysis was based on the production of 10 minutes of spontaneous speech by Moroccan, Tunisian and Lebanese subjects. Their findings demonstrate that rhythm variation across Arabic dialects is to a great extent correlated with the different types of syllabic structures observed in these dialects.

Why do we need a new corpus for studying Arabic rhythm?

- Previous Arabic rhythm studies were mainly relying on either read or spontaneous data. While our corpus will involve different speaking styles (read and spontaneous), which allows a better understanding of the nature of Arabic rhythm.
- A lot of researchers use an appropriate translation of "The North Wind and the Sun," (a standardized phonetic research text):
    - This doesn't guarantee that the translation would sound natural to the native speakers to read.
    - The translated text might not be phonetically balanced.

  Those problems of translated text will affect their reading which subsequently affects the speech rhythm measures. The current corpus will include originally Arabic Text, which will be easy for native speakers to read.

- Two issues regarding previous Arabic studies are the limited number of speakers and sentences, while the current study will overcome those problems and plan to include a reasonable number of materials.
- By following the same recording procedures of BonnTempo corpus, this corpus will help to study Arabic rhythm with a concrete comparison with other languages in BonnTempo corpus (German, English, French and Italian).
- The current corpus will help researchers to explore between- and within-speaker rhythmic variability among Arab speakers in the presence of different speech rates.
- There are huge variations between MSA and the Egyptian Colloquial Arabic (Kirchhoff and Vergyri, 2005); so adding both varieties will contribute to our understanding of Arabic rhythm in the MSA and dialectical form.

To investigate questions about Arabic rhythm and to place Egyptian Arabic within languages in general and stressed-timed ones, the current Arabic corpus has been built. Below, the corpus design including speakers, speaking styles, recording sessions, and recording set-up is being elaborated.

2 Corpus building & Recording

2.1 Speakers

The current corpus consists of recordings from 10 gender-balanced native Egyptian Arabic speakers (and is planned to include more). The participants were aged between 21 and 35 years (mean = 22.8, sd = 3.76). They were originally from the city of Alexandria (North of Egypt). Eligibility criteria required individuals to demonstrate little to no regional and social accent variability. All participants were recruited from the university environment and they didn't report any speech, language or hearing disorder.

2.2 Materials

The speech material of the present corpus consists of two speaking styles (read and spontaneous), which is captured using three tasks (see Table 1). The first style (read) replicates the methodology used in the creation of BTC. While for the spontaneous style, the participants were asked to speak freely for one or two minutes.

Type	Task	Duration
Read	Short story	~ 40 minutes
Spontaneous	Interview questions	~ 15 minutes
Spontaneous	Map (directions)	~ 15 minutes

Table 1: Speaking tasks for each speaker

The speech material in BonnTempo (BTC) currently consists only of read utterances but they are planning to include spontaneous speech in the future. The text is a short passage from a novel with 76 syllables in the German version. This text has been translated into the other languages under investigation by philologically educated native speakers of the target languages, Czech (93 syllables), English (77 syllables), French (93 syllables), Italian (106 syllables).

In the current corpus, we follow the same speech material structure of BTC (read speech). Speakers read a phonetically rich and balanced short story paragraph. The passage was subdivided roughly into 8 sentences with 178 phonological syllables. The total duration of utterances was around 4 minutes per speaker. A professional Arabic linguist manually added full diacritical marks to the written sentences. The reason for that is to avoid any ambiguity in pronunciation and enforce correct articulation. Sentences are phonetically rich (consist of the entire Arabic phonemic inventory) and balanced (having the same appearance in the language). The read part of this corpus consists of 400 tokens: 8 sentences X 10 speakers X 5 intended speech rate. The following Figure 1 describes the distribution of syllables in the corpus.

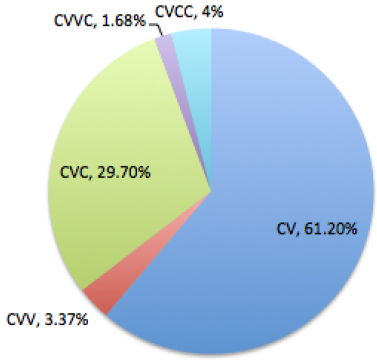

Figure 1: Syllables distribution of the read story

In addition to that, we also add a new recording material of spontaneous speech tasks; spontaneous speech data are usually preferred in linguistics analysis as they are closer to natural speech. The speakers were asked to answer two interview questions about their daily life and also to describe the directions (for details see the procedure section).

2.3 Recording procedure

The following section details the corpus procedures and includes descriptions of the recording tasks and the setup.

2.3.1 Speaking tasks

As mentioned above, this corpus consists of two speaking styles (read and spontaneous), which is captured using three tasks.

Task1: Read a short story

During the recording process for the reading task, speakers were given the text of the story first to familiarize themselves with the text by reading it aloud before the recording. Speakers were allowed to practice the text as many times as they wanted before the actual recording started. For the first recording, they were asked

to read the text in a way they considered 'normal reading' (no). After that they were asked to read the same text at different intended speech rates (slow, slower, fast and fastest possible): Firstly speakers were asked to read it slowly (s1) and then even more slowly (s2). Following the recordings at slow rates, they were required to read the text fast (f1) and then they consecutively had to increase their reading speed to the fastest as they could (f2).

Task 2: Interview questions

The second task is interview questions. To capture spontaneous speech, speakers were asked to talk freely for more than one minute (or around 8 sentences) about their daily life and their studies. As they are all university students from the same department, we assume that their answers will share similar content.

Task 3: Map direction

The third task is map direction description, which contains spontaneous speech generated by using a map as a visual stimulus in order to encourage the elicitation of more speech. The speakers were asked to describe the directions to go to the university from a famous nearby location.

2.3.2 Recording set-up and sessions

Recordings were carried out in the soundproof room at Alexandria University with a large membrane condenser microphone directly on PC in .wav file format. The recordings have a 44.1 KHz sampling rate and 16-bit quantization. The microphone was located on approximately 40 cm distance from the speaker's lips. Speakers were asked to read the short story and produce the spontaneous speech in two non-contemporaneous sessions. Due to the importance of accounting for between-speaker variability and based on the criteria of the non-contemporaneity of recording sessions for each speaker in the protocol for the collection of databases of recordings for forensic-voice-comparison research and practice, all the speakers were recorded twice, on two recording sessions taking place on different days. Recording sessions were separated by a time-lapse of one to two weeks.

3 Corpus analysis and annotation

Speech tokens were analyzed using Praat (version 5.3.78)(Boersma and Weenink 2019). Firstly, segments on- and offsets were labeled manually using Praat's annotation function. The utterances were phonemically transcribed with IPA symbols. We used the waveform, spectrogram and auditory discrimination cues in determining phoneme boundaries. Consonantal and vocalic intervals are, next to the syllables, the most central and most often used units of speech for rhythmic measurements in speech (Dellwo, 2010). For this reason, we annotated our data based on the aforementioned units. CV intervals were created automatically using an automatic script *CV Creator Tier*. A C-interval consists of one or more consonants preceded and followed by a vowel or by a pause whereas a V-interval consists of one or more vowels (or vocalic segments like diphthongs, triphthongs, etc.) preceded and followed by a consonant or by a pause (Dellwo, 2010). CV intervals comprise three tiers, (a) a tier containing consonantal and vocalic segments, (b) a tier containing consonantal and vocalic intervals with each interval containing the number of underlying consonantal and vocalic segments respectively and (c) a tier containing consonantal and vocalic intervals. The syllable tier was also labeled manually by trained phoneticians by following Arabic phonotactic rules for syllabification (see Figure 3)

All five tempo versions (s2, s1, no, f1, f2) of each speaker have been saved in wav format in one file each (see Figure 2). The file names contain information about the native language of the speaker in capital letters (e.g. 'Ar' for Arabic), speaker number, and the speaking tempo (e.g. 'no' for normal). Language, speaker's number, and tempo information are separated by an underscore (e.g.: Ar_01_no.wav = Arabic native speaker, number 1, intending to read in normal tempo).

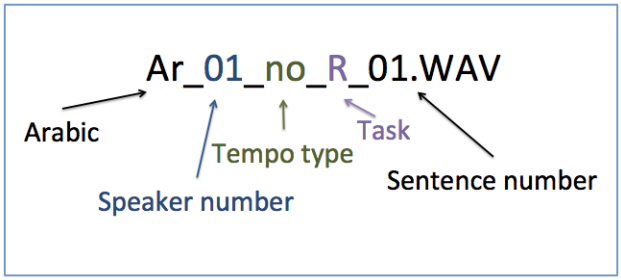

Figure 2: File naming convention

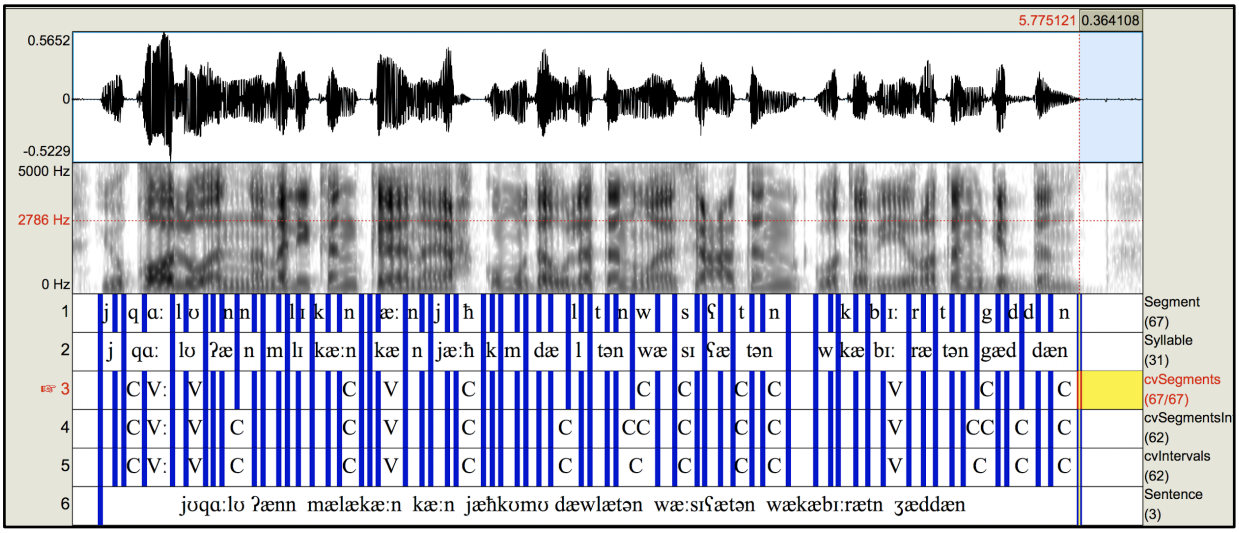

Figure 3: Example of corpus annotation (Praat TextGrid) of one sentence "Once up on a time there was a king who ruled a wide and huge kingdom"

The entire recorded corpus was transcribed orthographically because of many reasons; first, it provides further researchers with a simple symbolic representation of the recorded data. With this representation, it is easy to navigate through the corpus. Secondly, the orthographic transcription formed the basis for all other transcriptions and annotations. Thirdly with regard to rhythmic measurers, the extraction of interval duration from speech relies mostly on manual inspection of waveforms and spectrograms and is therefore subject to the vagaries of individual researchers, who may use different criteria or apply common criteria idiosyncratically. The use of automatic methods of extraction is a clear first step for maintaining consistency.

The annotation work for each speaker has been saved in Praat label files of the type 'TextGrid'. For each wav file there exists one TextGrid file with the same file name but respective extension (e.g. Ar_01_no_R_01.TextGrid).

4 Conclusion

In this paper, we presented the development of a homogeneous database for Arabic in Egyptian dialect, specifically designed for FVC tasks. We collected our corpus based on a) the protocol for the collection of databases of recordings for forensic-voice-comparison research and practice and, b) BonnTempo corpus. We plan to conduct research on between- and within-speaker variability of supra-segmental acoustic parameters in our database. We also aim to study the degree to which acoustic cues vary across different speaking styles and what affects this variability has on speaker identification. As well as dealing with forensic issues, which is our primary goal in this project, the described database paves the way for addressing a number of theoretical and practical issues in acoustic phonetics of the Arabic language. This database is supposed to be developed further and we aim to increase the number of speakers and speaking styles in the future.

As there are many varieties of colloquial Arabic, some are mutually intelligible, while others are not and the larger the physical distance between the dialects, the more a difference appears among them (Hetzron, 1997). For the future extension of the corpus, other Arabic dialects like Morocco and Jordan Arabic are planned to be recorded with the same procedures for Arabic dialects comparison.

5 Bibliographical References

Alghamdi M., Alhamid A., and Aldasuqi M.,(2003). Database of Arabic Sounds: Sentences, Technical Report, Saudi Arabia.

Al-Sobh, M., Abu-Melhim, A., &Bani-Hani, N. (2015). Diglossia as a result of language variation in Arabic: Possible solutions in light of language planning. Journal of Language Teaching and Research, 6(2),274- .27

Al-Suwaiyan L. A. (2018). Diglossia in the Arabic Language, International Journal of Language and Linguistics,Vol. 5, No. 3, September 2018, doi:10.30845/ijll.v5n3p22

Altuwaim, Y. A. Alotaibi and S. Selouani (2014). Investigation into the speech rhythm of two Saudi dialects using the SAAVB corpus. 2014 6th International Symposium on Communications, Control and Signal Processing (ISCCSP), Athens, 2014, pp. 632-635. doi: 10.1109/ISCCSP.2014.6877954

Bateson, M. C. (2003). Arabic Language Handbook. Washington, Georgetown University Press.

Boersma, Paul & Weenink, David (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1.06, retrieved 8 November 2019 from http://www.praat.org/

Dellwo, V., & Fourcin, A. (2013). Rhythmic characteristics of voice between and within languages. TRANEL - Travaux neuchâtelois de linguistique, 59, 87–107.

Dellwo, V., Leemann, A., and Kolly, M.-J. (2015). Rhythmic variability between speakers: articulatory, prosodic, and linguistic factors, Journal of the Acoustical Society of America 137: 1513-1528.

Dellwo, V., Steiner, I., Aschenberner, B., Dankovičová, J., and Wagner, P. (2004). The BonnTempo-Corpus and BonnTempo-Tools: A database for the study of

speech rhythm and rate. in Proceedings of the 8th ICSLP, Jeju Island, Korea.
Droua-Hamdani, S.-A. Selouani, M. Boudraa, W. Cichocki, (2010). Algerian arabic rhythm classification., in: ExLing, pp. 37–40.
Enzinger E. and Morrison G. S. (2012). The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems. in *the 14th Australasian International Conference on Speech Science and Technology, Sydney, Australia, Proceedings*, pp. 137-140.
Ghazali, S./R. Hamdi, R./M. Barkat, M. (2002). Speech rhythm variation in Arabic dialects. – In: Bernard Bel/I. Marlin (eds.), Proceedings of the Speech Prosody 2002 Conference, 11-13 April 2002, Aix-en-Provence: Laboratoire Parole et Langage, 127-132.
Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In Papers in Laboratory Phonology VII (Eds. Gussenhoven, E. & Low, E. L.), Berlin: Mouton de Gruyter, 515–546.
Grabe, E., Low, E.L. (2003). Durational variability in speech and the rhythm class hypothesis. Papers in laboratory phonology 7, 515-546.
Hetzron, R. (1997). Classical Arabic.The Semitic languages. New York, NY: Routledge.
Kinoshita, K. (2001). *Testing realistic forensic speaker identification in Japanese: A likelihood ratio based approach using formants*. Ph.D. dissertation, Australian National University.
Kirchhoff, K. and Vergyri, D. (2005). Cross-dialectal data sharing for acoustic modeling in arabic speech recognition. *Speech Communication*, 46(1):37–51.
Hamdi, R., Ghazali, S., & Barkat-Defradas, M. (2005). Syllable Structure in Spoken Arabic: A comparative investigation. Interspeech, 2245– 2248.
He, L. and Dellwo, V. (2016). The role of syllable intensity in between-speaker rhythmic variability. *International Journal of Speech, Language and the Law 23*(2): 243–273. https://doi.org/10.1558/ijsll.v23i2.30345
Jang, T-Y. (2009). Automatic assessment of non-native prosody using rhythm metrics: Focusing on Korean speakers' English pronunciation. In SFU Working Papers in Linguistics Vol. 2 . Simon Fraser University, Vancouver, Canada.
Leemann, M.-J. Kolly, and V. Dellwo (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. Forensic Sci. Int. 238, 59–67.
Marcus, S. M. (1981). Acoustic determinants of perceptual center (P-center) location. Perception & Psychophysics, 30, 247–256.
Morrison, G. S., Rose, P., and Zhang, C. (2012). "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice," Aus. J. of Forensic Sci. doi:10.1080/00450618.2011.630412.
Nolan, Francis and Eva Liina Asu. (2009). The Pairwise Variability Index and Coexisting Rhythms in Language. Phonetica 66 (1-2),pp. 64-77.
O'Rourk, E. (2008). Speech rhythm variation in dialects of Spanish: Applying the Pairewise Variability Index and variation Coefficients to Peruvian Spanish. Speech Prosody, 6-9 May, Brazil.
Ramus, F., Nespor, M., Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. Cognition 72,1-28.
Rose, P. (2002) Forensic Speaker Identification. New York: Taylor & Francis. https://doi. org/10.1201/9780203166369
Watson, JCE (2011). Word stress in Arabic. In: The Blackwell companion to phonology. Wiley-Blackwell, Oxford, 2990-3019 (p. 2991)