Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2013) 21: 1426 – 1439
© TÜBİTAK
doi:10.3906/elk-1112-29

Review of distinctive phonetic features and the Arabic share in related modern research

Yousef ALOTAIBI,* **Ali MEFTAH**
College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

Received: 07.12.2011 • **Accepted:** 31.05.2012 • **Published Online:** 12.08.2013 • **Printed:** 06.09.2013

Abstract: Most research in the field of digital speech technology has traditionally been conducted in only a few languages, such as English, French, Spanish, or Chinese. Numerous studies using distinctive phonetic features (DPFs) with different techniques and algorithms have been carried out during the last 3 decades, mainly in English, Japanese, and other languages of industrialized countries. DPF elements are based on a technique used by linguists and digital speech and language experts to distinguish between different phones by considering the lowest level of actual features during phonation. These studies have investigated the best performances, outcomes, and theories, especially those regarding digital speech recognition. The aim of this paper is to present the background of DPF theories and the usefulness thereof for digital speech and language processing. In addition, we highlight the background of Arabic language phonology compared to 2 well-known languages to enhance the current knowledge about this narrow language discipline. Finally, this work reviews the research dealing with DPF strategies for digital speech and language processing using computing and engineering techniques and theories. Based on the literature search conducted for this paper, we conclude that although the Arabic language is a very important and old Semitic language, hitherto it has suffered from a lack of modern research resources and theories on DPF elements.

Key words: Arabic, DPF, speech, MSA, ASR, neural network

1. Introduction

1.1. Arabic language background

Arabic is one of the world's major languages. It is the fifth most widely spoken language in the world and is second in terms of the number of speakers, with over 250 million Arabic speakers, of whom roughly 195 million are first language speakers and 55 million are second language speakers [1]. To begin the study on distinctive phonetic feature (DPF) elements for the Arabic language, many points about the nature of this language must be taken into account. The Arabic language has 3 forms: classical Arabic, modern standard Arabic (MSA), and colloquial Arabic. Classical Arabic is the language of the Quran, the religious instruction in Islam, and of the great writers and poets. MSA (or *Al-fus*$^{?}$*ha*) is one of the Arabic dialects and is the form of the Arabic language that is taught in schools and used in most radio and television broadcasts, formal talks, and the majority of the printed matter in the Arab world, including books. Colloquial Arabic (or *al-ammiyya*) is the form of Arabic that is used in everyday oral communication.

Arabic dialects vary in many dimensions, but primarily with respect to geography and social factors. According to geographical linguistics, the Arab world can be divided in many different ways. Given below

[*] Correspondence: yaalotaibi@ksu.edu.sa

is only one of those covering the main Arabic dialects. Gulf Arabic includes the dialects of Kuwait, Saudi Arabia, Bahrain, Qatar, United Arab Emirates, and Oman, while Iraqi Arabic is the dialect of Iraq. In some dialect classifications, however, Iraqi Arabic is considered a subdialect of Gulf Arabic. Levantine Arabic includes the dialects of Lebanon, Syria, Jordan, and Palestine. Egyptian Arabic covers the dialects of the Nile valley, including Egypt and Sudan. Maghreb Arabic covers the dialects of Morocco, Algeria, Tunisia, and Mauritania. Libya is sometimes included in this class. Yemenite Arabic is often considered a class on its own. Socially, it is common to distinguish 3 subdialects within each dialect region: city dwellers, peasants/farmers, and Bedouins. The 3 levels are often associated with a 'social hierarchy' from the rich, settled city-dwellers, down to Bedouins. Different social associations exist, as is common in many other languages around the world [2].

1.2. Arabic, English, and Japanese phonology

Most of the research on speech processing has focused on the languages of industrial countries, such as English and Japanese. Moreover, most of the resources and supporting speech tools and products have been implemented using English and Japanese, together with the languages of other industrialized countries. Due to the familiarity of the literature, research, and speech tools for English and Japanese, we briefly compare Arabic with these 2 well-known languages.

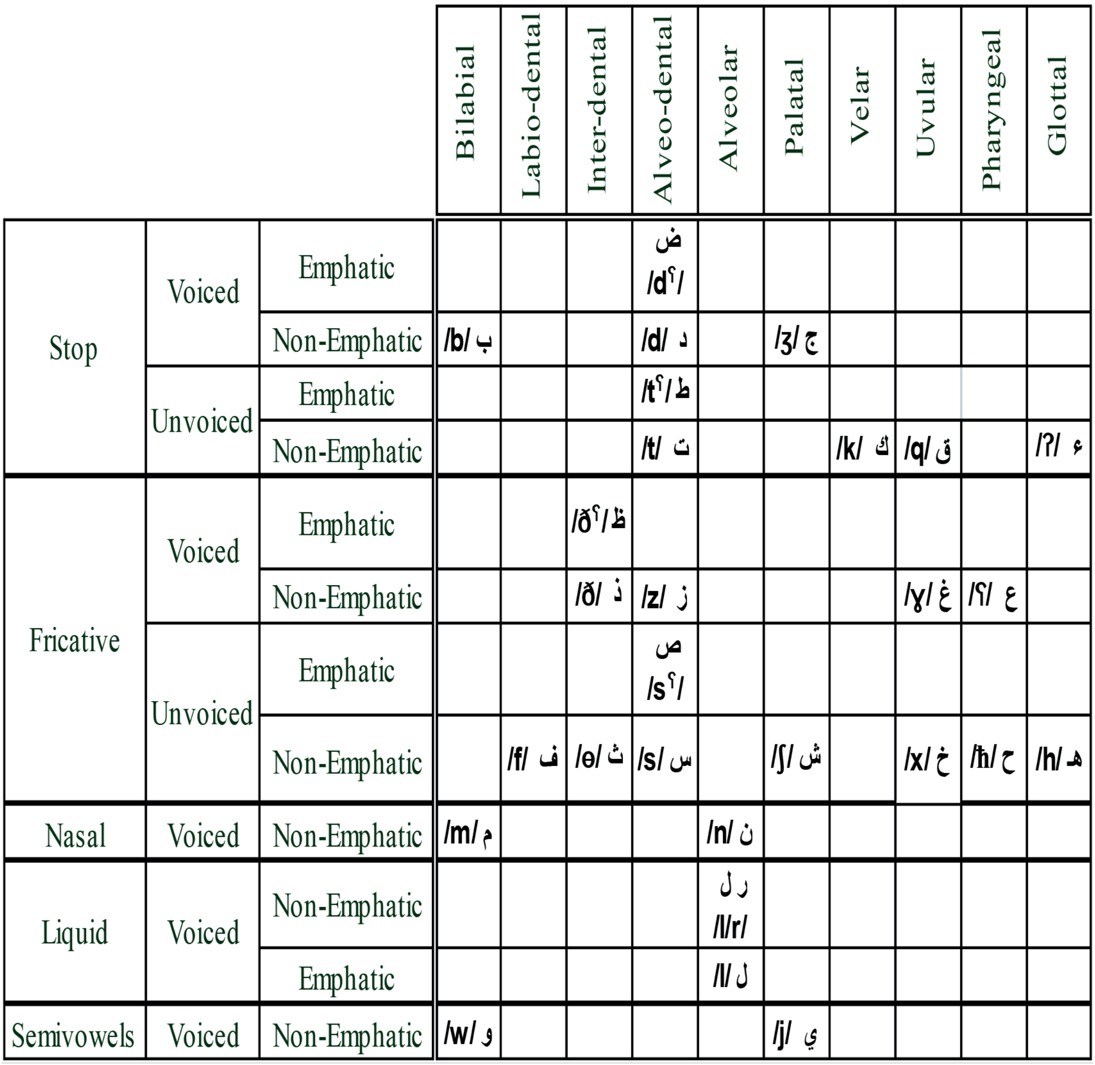
To be specific our aim is to investigate some of the phonological characteristics of Arabic, English, and Japanese by comparing the 3 languages with respect to their phonological level, specifically by investigating the vowels, consonants, and syllables. The differences and similarities are highlighted to obtain a clear background of these 3 important languages. Vowels and consonants are phonemes that can be defined as the smallest part of speech that designates a variation in the meaning of a spoken valid word of the language [3]. We intend to search the phoneme inventory of these 3 languages.

MSA Arabic has 6 vowels and 2 diphthongs. The 6 vowels are /a/, /i/, /u/, /aa/, /ii/, and /uu/, where the former 3 are short vowels and the latter 3 are the corresponding longer versions of the short vowels. On the other hand, the 2 diphthongs are /ae/ and /ao/ [3–5]. As a result, the duration of the vowel sounds is phonemic in the Arabic language, and this is one of the major distinctions of Arabic compared to English and Japanese. Each short Arabic vowel is phonetically identical to its long counterpart (i.e. the only difference is the duration thereof) [5]. Arabic dialects may have different and additional vowels; for instance, the Levantine dialect has at least 2 extra diphthongs, /aj/ and /aw/. Similarly, the Egyptian dialect has additional vowels [6].

Comparing English and Japanese, we see some significant differences in the following 2 areas: first, the number of vowels, and second, the tense/lax distinctions. In the English vowel system, there are more than 13 different vowels (depending on the American and British dialects) and these include several diphthongs. On the other hand, Japanese has only 5 vowels, which are common to most languages, including Arabic and English [7,8]. Another characteristic that differentiates the English vowel system specifically from the Japanese vowel system is whether there is a distinction between the lax and tense vowels in either of the 2 systems. The differentiation between the tense and lax vowels is made according to how much muscle tension or movement in the mouth is involved in producing the vowels [8]. Vowels produced with extra muscle tension are called tense vowels, while those produced without that much tension are called lax vowels. For example, /i/ as in the English word "eat" is categorized as a tense vowel as the lips are spread (muscular tension in the mouth) and the tongue moves toward the root of the mouth. To be specific, the tense/lax vowel pairs in English, such as /i/ vs. /I/, /e/ vs. /ε /, and /u/ vs. /U/, do not exist in the 5-vowel Japanese system, because the tense/lax differentiation is not phonemic [7,8].

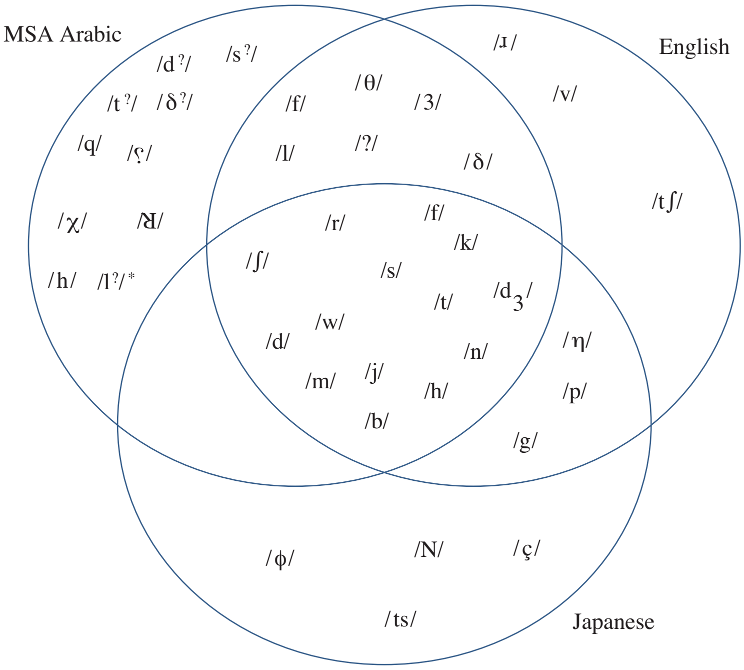
Regarding consonants, MSA Arabic contains 28 consonants, varying between stops, fricatives, nasals, and liquids. Moreover, the Arabic consonant system comprises 2 distinctive classes, known as pharyngeal and emphatic phonemes. These 2 classes cannot be found in either English or Japanese, but can be found in other Semitic languages such as Hebrew [4]. Regarding the consonant systems, there are clear differences in the consonantal distributions between Arabic, Japanese, and English. One of the differences is the lack of affricates in Japanese, which is not the case in Arabic and English. Moreover, regarding the point of articulation, there is a variety of fricatives and affricates, which are much more widely distributed in Arabic and English than in Japanese. Specifically, /v/, /θ /, /ð/, / ʒ /, and /dʒ/ are found in the Japanese consonantal system, but some of these phonemes exist in Arabic and English [7]. Another difference in the consonantal system between Japanese and English is that there are some consonants found in the consonant inventory of Japanese, but not in that of English, such as the voiceless bilabial fricative /Φ/ and voiceless palatal fricative /Ç/ [8]. In addition, Japanese has a liquid consonant that does not correspond exactly to the English liquid /r/ or /l/, but rather

Table 1. Arabic and English consonants [3–5,9].



			Bilabial	Labio-dental	Inter-dental	Alveo-dental	Alveolar	Palatal	Velar	Uvular	Pharyngeal	Glottal
Stop	Voiced	Emphatic				ض /dˤ/						
		Non-Emphatic	/b/ ب			/d/ د		/ʒ/ ج				
	Unvoiced	Emphatic				/tˤ/ ط						
		Non-Emphatic				/t/ ت			/k/ ك	/q/ ق		/ʔ/ ء
Fricative	Voiced	Emphatic			/ðˤ/ ظ							
		Non-Emphatic			/ð/ ذ	/z/ ز				/ɣ/ غ	/ʕ/ ع	
	Unvoiced	Emphatic				ص /sˤ/						
		Non-Emphatic		/f/ ف	/ɵ/ ث	/s/ س		/ʃ/ ش		/x/ خ	/ħ/ ح	/h/ هـ
Nasal	Voiced	Non-Emphatic	/m/ م				/n/ ن					
Liquid	Voiced	Non-Emphatic					ر ل /l/r/					
		Emphatic					/l/ ل					
Semivowels	Voiced	Non-Emphatic	/w/ و					/j/ ي				

is considered to be a sound in-between that of the English /r/ and /l/. The exact articulation point is not specified for the Japanese /r/ sound. Hence, the most characteristic difference between the Japanese and English consonantal systems lies not in the number of consonants found in each of the 2 languages, but rather in the unique distribution patterns of the consonants in each language. Table 1 displays the full features of consonants of MSA Arabic and English in a way that facilitates comparison. In addition, using set methodology, Figure 1 displays the common phonemes between any pairs of the 3 investigated languages, as well as the common set for all 3 languages [7] The contents in Figure 1 were collected from different well-known references dealing with these 3 important languages [3–5,9–11].



*Exist only in one valid MSA Arabic word

Figure 1. Arabic, Japanese, and English consonants [3–5,9–11].

Regarding the syllables, an MSA Arabic syllable must contain at least 1 vowel. Moreover, Arabic vowels cannot appear as the first part of a syllable, but can occur either between 2 consonants or at the end in a syllable or word. Arabic syllables can be classified as short or long. The allowed syllables in the Arabic language are: CV, CVC, and CVCC where V indicates a (long or short) vowel and C indicates a consonant. The CV type is a short one, while all of the others are long. Syllables can also be classified as open or closed. An open syllable ends with a vowel, whereas a closed syllable ends with a consonant. For Arabic, a vowel always forms a syllable nucleus, and there are as many syllables in a word as there are vowels [5,12].

On the other hand, English allows a wide variety of syllable types including both open and closed syllables: CV (open syllable), CVC, CCVC, CCVCC, and CCCVCC (closed syllable).

In Japanese, the allowed syllable types seem to be restricted to open syllables only. The fact that Japanese words of more than 1 syllable always follow the CV-CV-CV syllable sequence clearly shows the significant characteristics of Japanese syllables, which differ from those in English [7].

The main differences here are: first, Japanese does not allow a word to end with a consonant, which is exactly the case in Arabic, and second, Japanese does not permit both initial and final consonant clusters (i.e. a CCVCC syllable). Thus, in general, Arabic and English have a wider range of syllable types than Japanese, and they also allow the occurrence of consonant clusters, both in the initial and final positions in a word.

There is a distinction between Arabic and English in this regard, in that Arabic syllables start with a single consonant only followed by a vowel, whereas an English syllable can start with 1, 2, or even 3 consonants [5]. It should be noted, however, that although English permits initial and final consonant clusters, there are some restrictions on the possible combinations of consonants when realized in consonant clusters [7].

2. Arabic speech, language, and DPF

2.1. DPF elements

Jacobson and Chomsky were among the first scientists to work on DPF elements in English and additional languages in the last century [9,13]. DPF elements are a way of representing phoneme phonation by specifying the manner of articulation (vocalic, consonantal, continuant . . .), tongue position (high, front, end . . .), etc. and can separately uniquely identify each phoneme. According to various researchers' definitions, DPF elements are a compact set of articulatory and acoustic "gestures", combinations of which can codify meaningful similarities and dissimilarities between all sounds [9]. Each phoneme can then be represented by its values in the distinctive feature space, and the differences between the phones can be described simply and succinctly in the same space [9]. In other words, together with other definitions by various language experts, speech sounds can be described using a DPF representation by identifying a set of physiological actions or states that serve to distinguish speech sounds from one another. Any language phonemes are viewed as a shorthand notation for a set of DPF elements, which describe the operations of the articulators required to produce the distinctive aspects of a speech sound [14]. The DPF is related to another term, that is, the articulatory features defined as the group of properties of a speech sound based on its voicing or on its place or manner of articulation in the vocal tract from a phonetic point of view [13]. In other words, DPF depends on linguistic and phonemic observation, whereas articulatory features depend on the phonetic characteristics of the actual vocal tract while vocalizing the specific language phonemes. Linguists can easily identify a phone using the values of the DPF elements. This can be done by identifying a set of physiological actions or states, including high, low, anterior, back, coronal, plosive, continuant, fricative, nasal, voiced, and semivowel, which help to distinguish the speech sounds from one another [14,15]. Based on another variation in the definition, phonemes are viewed as the shorthand notation for a set of features describing the operations of the articulators required to vocalize the distinctive aspects of a speech sound. To give an example, the phonemes "p" and "b" are produced in ways that differ only in the state of the vocal folds; "p" is produced without a vibration (unvoiced), while "b" requires a vibration of the vocal cords (voiced). In the distinctive feature representation, only the feature "voice" differs for these 2 sounds [14]. For the Arabic language, Table 2 shows the common DPF elements as agreed upon by Arabic linguistics and researchers.

To achieve automatic speech recognition (ASR) at the highest possible levels of performance, we have to ensure efficient use of all of the contextual and phonetic information. The specific phones that are used in any instance depend on contextual variables such as the speaking rate. On a short time scale, such as the average length of a phone, limitations on the rate of change of the vocal tract cause a blurring of the acoustic features that is known as the coarticulation effect. For longer time scales there are many contextual variables that vary only slightly (e.g., the degree and spectral characteristics of the background noise and channel distortion) and speaker dependent characteristics (e.g., vocal tract length, speaking rate, and dialect) [16]. Compared to other major languages in the world, the Arabic language generally suffers from a lack of research initiatives and modern research resources, especially on the topic of DPF and applications in digital speech processing.

Table 2. MSA Arabic DPF elements [9,13,17].

Arabic Writing	IpA Symbol	affricative	alveodental	alveoplatal	aspirated	bilabial	consonant	continuant	emphatic	fricative	glottal	interdental	labiodentals	labiovelar	lateral	nasal	palatal	pharyngeal	plosive	rounded	semivowel	trill	unvoiced	uvular	velar	voiced	vowel	anterior	coronal	high	SIL
	sil	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
ب	b	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	+	-	-	-
ت	t	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	+	+	-	-
ث	θ	-	-	-	-	-	+	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	+	-	-
ج	ʒ	+	-	+	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
ح	h	-	-	-	-	-	+	+	-	+	-	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-
خ	x	-	-	-	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-
د	d	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	+	+	-	-
ذ	ð	-	-	-	-	-	+	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	-	-
ر	r	-	+	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	+	+	-	-
ز	z	-	+	-	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	-	-
س	s	-	+	-	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	+	-	-
ش	ʃ	-	-	+	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	+	-	-
ص	s^{ʔ}	-	+	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	+	+	-
ض	d^{ʔ}	-	+	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	+	+	+	-
ط	t^{ʔ}	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	+	+	+	-
ظ	ð^{ʔ}	-	-	-	-	-	+	+	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	+	-
ع	ʔ	-	-	-	-	-	+	+	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-
غ	γ	-	-	-	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	+	-
ف	f	-	-	-	-	-	+	+	-	+	-	-	+	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-
ق	q	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	+	-
ك	k	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	+	-	-	-	-	-	-
ل	l	-	+	-	-	-	+	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	+	+	-	-
م	m	-	-	-	-	+	+	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-
ن	n	-	+	-	-	-	+	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-	+	+	-	-
هـ	h	-	-	-	-	-	+	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
و	w	-	-	-	-	+	-	+	-	-	-	-	-	+	-	-	-	-	-	+	+	-	-	-	-	+	-	+	-	-	-
ي	j	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	+	-	-	-	-	-
ـا	aa	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	+	-	-
ـو	uu	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	+	+	+	-	-	-
ـي	ii	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-

In this paper, we review several works related to the DPF elements in 3 subsections: DPF in the Arabic language, artificial neural network (ANN) techniques used in DPF, and other techniques considered for DPF. For each section, the papers are ordered from oldest to newest.

2.2. Studies on Arabic DPF

The Arabic language has unique characteristics; for example, pharyngeal and emphatic, together with various kinds of allophone germination. Table 2 gives the Arabic DPF elements that exist in its main dialect namely MSA. These phonemes and their assigned features can vary in different Arabic regional and local dialects. Moreover, the Arabic language has more lexical stress systems than any other language, but regrettably, in our investigation of the Arabic language and DPF, we have found only a small number of good studies and papers on this subject.

In this context, Selouani et al. [18] worked on spotting Arabic phonetic features using a modular connectionist architecture and rule-based system. They presented the results of their experiments in complex Arabic phonetic feature identification using a rule-based system and modular connectionist architectures. The aim of their study was to use the Arabic language to test the ability of automatic systems operating a 'blind' classification for detecting aspects as subtle as germination, and emphasis and relevant extension of vowels. They used 2 techniques. The first operates in the field of analytic approaches and aims to implement a relevant system for automatic segmentation and labeling through the use of finite state networks. The second deals with a set of a simplified version of subneural networks. Two types of architectures, namely, serial and parallel architectures of subneural networks, were investigated. A comparison between the 2 identification strategies was executed using stimuli uttered by Algerian native speakers. Based on the results, the authors concluded that for the detection of complex phonetic features such as the phonological duration, i.e. long vowels and their germination, the rule-based system is more promising. In contrast, when a rough discrimination is solicited, neural networks are more adaptable and the parallel architecture of subneural networks is the most reliable system. They claimed that their proposed connectionist mixture of experts is advantageous in that it simplifies learning because the binary discrimination does not need a large number of cycles. Moreover, generalization of the identification of other features such as the speaker's sex, voiced-unvoiced markers, etc. may constitute a simple, yet powerful, way of improving ASR systems.

Selouani et al. [19] also worked on Arabic phonetic feature recognition using modular connectionist architectures. They proposed an approach for reliably identifying complex Arabic phonemes in continuous speech using a mixture of artificial neural experts. Their objective was to test the ability of autoregressive time delay neural networks to detect Arabic complex phonemes. The authors claimed that the parallel and serial structures of autoregressive time delay neural networks surpass the monolithic configuration and the parallel disposition constitutes the most reliable system. Again, as in their previous work, they claimed that the proposed mixture of the neural experts approach is advantageous since it facilitates learning because binary discrimination does not need a large number of cycles. Generalization to the identification of other features, such as the speaker's sex and prosodic features may constitute a simple yet powerful way of improving the performance of ASR systems.

3. Used techniques with DPF

3.1. ANN techniques

There is a strong relation between neural networks and speech DPF features. A human being's brain can easily identify the different distinctive features in speech; hence, brains can classify language phonemes easily and accurately. For example, the brain can distinguish between the phonemes /s/ and /z/, where the only difference is just the voicing, which is missing in the former but is present in the latter phoneme. Moreover, the human brain can be best emulated by ANN technology in our computational intelligent artificial system nowadays. Thus linguists and computer scientists would agree on the fact that ANNs may be the best choice in order to identify the speech units' features.

In this section, we review different studies by several researchers working on neural network techniques along with DPF technology and digital speech processing to achieve the best outcomes, and put forth theory and guidance for future researchers in this interesting, important, and relevant research subject.

King et al. [20] focused on the detection of phonological features in continuous speech using neural networks. A description of the techniques for detecting the phonological features in continuous speech was included in their work. They reported work on speech recognition architectures based on phonological features rather than phones. Their experiments focused on 3 phonological feature systems: the sound pattern of the English system, which uses binary features; a multivalued feature system using traditional phonetic categories such as manner and place; and, finally, government phonology, which uses a set of structured primes. In all of these experiments the authors used recurrent neural networks to perform the feature detection. All of the experiments were carried out on the well-known English TIMIT speech corpus. The authors asserted that the networks performed well in all cases, with the average accuracy for a single feature ranging from 86% to 93%.

In another study, Launay et al. [21] focused on knowledge-based features for hidden Markov model (HMM)-based large vocabulary automatic speech recognition. They described an attempt to build a large vocabulary ASR system using distinctive features by replacing features based on their short-term spectra, such as mel-frequency cepstral coefficients (MFCC), with features that explicitly represent some of the distinctive features of the speech signal. The authors engineered an approach whereby neural networks are trained to map short-term spectral features to the posterior probability of some distinctive features. Experimental results on the Wall Street Journal task showed that such a system does not outperform the MFCC-based system, although it generated very different error patterns. They claimed that they were able to obtain reductions in the word error rates of 19% and 10% on the 5 K and 20 K tasks, respectively, over their best MFCC-based system. They suggested that their proposed approach could be a very promising way of incorporating speech knowledge into large-scale ASR systems.

Fukuda et al. [10] investigated DPF element extraction for robust speech recognition. They described an attempt to extract DPF elements that represent articulatory gestures in linguistic theory using a multilayer neural network and to apply the DPF elements to noise-robust speech recognition. In the DPF element extraction stage, after a speech signal has been converted to acoustic features, it is composed of local features. The authors mapped the acoustic parameters to the DPF elements using a multilayer neural network with context-dependent output units. The local features showed better performance than the MFCC as input for the multilayer neural network. The results of the proposed DPF without the conventional MFCC parameter are almost the same as the standard MFCC-based feature parameter in HMM-based isolated spoken word recognition experiments with clean speech. Moreover, this could significantly reduce the effect of high-level additive noise, particularly the ring tone of a mobile phone.

Fukuda et al. [22] focused on the idea of the canonicalization of the feature parameters for automatic speech recognition. They showed that the acoustic models of an HMM-based classifier include various types of hidden variables such as sex, speaking rate, and acoustic environment. They stated that a robust ASR system could be realized if there was a canonicalization process to reduce the influence of the hidden variables from the acoustic models. In their paper, they described the configuration of a canonicalization process targeting sex as a hidden variable. The authors proposed a canonicalization process composed of multiple DPF extractors corresponding to the hidden variable and a DPF selector, to compare the distance between the input DPF and the acoustic models. In the DPF extraction stage, an input sequence of the acoustic feature vectors was mapped onto 3 DPF spaces corresponding to male, female, and neutral voices using 3 multilayer neural networks. The paper concluded that the proposed canonicalization process could reduce the influence of sex as a hidden variable from the acoustic models.

In addition to the above, Huda et al. [16] worked on DPF-based phonetic segmentation using recurrent neural networks. They emphasized that possessing an ASR system at the highest possible level of performance implies the efficient use of all of the contextual information. The specific phones that are used in any instance depend on the contextual variables, such as the speaking rate. In their work, a 2-stage system of recurrent neural networks and a multilayer neural network was introduced to obtain better contextual information, and hence better phonetic segments. The experiments on several methods show that a better contextual effect can be obtained using a combination of a recurrent neural network and a multilayer neural network, rather than using only a multilayer neural network.

Huda et al. [23] also worked on DPF-based phone segmentation using a 2-stage multilayer neural network. In their paper, they introduced a DPF-based feature extraction using a 2-stage multilayer neural network, where the first stage maps the continuous acoustic features, namely the local features onto discrete DPF patterns, and the second stage constrains the DPF context or dynamics in an utterance. The experiments were carried out using Japanese triphthong data. The authors asserted that the proposed DPF-based feature extractor provides good segmentation and high recognition rates with a reduced mixture-set of HMMs by resolving coarticulation.

Again, but with a slight variation with respect to the type of corpus used, Huda et al. [15] worked on DPF-based phone segmentation using hybrid neural networks. In their work, they introduced DPF-based feature extraction using a 2-stage neural network system consisting of a recurrent neural network in the first stage and a multilayer neural network in the second stage. The recurrent neural network maps continuous acoustic features, i.e. local features, onto discrete DPF patterns, while the multilayer neural network constrains the DPF context or dynamics in an utterance. The experiments were carried out using Japanese newspaper article sentences, and continuous utterances containing both vowels and consonants. Again the authors argued that the proposed DPF-based feature extractor provides good segmentation and high recognition rates with a reduced mixture-set of HMMs by resolving the coarticulation effect.

For a second time, Huda et al. [24] investigated phoneme recognition based on hybrid neural networks with the inhibition/enhancement of DPF trajectories. They presented a novel DPF extraction method that incorporates inhibition/enhancement functionalities by discriminating the DPF dynamic patterns of the trajectories as relevant or not. The proposed algorithm, which enhances convextype patterns and inhibits concave-type patterns, was implemented in a phoneme recognizer and evaluated. Their recognizer consists of 2 stages. The first stage extracts 45 dimensional DPF vectors from local features of input speech using a hybrid neural network and incorporates an inhibition/enhancement network to obtain modified DPF patterns. The second stage orthogonalizes the DPF vectors, and then feeds these to an HMM-based classifier. They concluded that the proposed phoneme recognizer significantly improves the accuracy of the phoneme recognition with fewer mixture components by resolving coarticulation effects. Finally, Yu et al. [25] worked on boosting the attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition. They achieved high accuracies for both phonological attribute detection and phone estimation by using deep neural networks.

3.2. Other techniques

In this section, we consider a variety of different studies that used DPF with techniques other than neural networks. Initially, Nitta [11] worked on feature extraction for speech recognition based on orthogonal acoustic-feature planes along with linear discriminant analysis. He described an attempt to extract multiple topological structures, hidden in time-spectrum patterns, using multiple mapping operators, and to incorporate the operators into the feature extractor of a speech recognition system. His design methodology for mapping the operators in the feature extractor was created by observing the orthogonal basis of speech and modeling it. The proposed method based on multiple acoustic-feature planes along with linear discriminant analysis showed significant improvements compared with the conventional time-spectrum method, as well as the Karhunen–Loève transform, time-spectrum, and linear discriminant analysis in the experiments using the Japanese Cv-set speech database. Furthermore, the proposed method maintains accuracy in the range of small feature dimensions.

Eide [14] worked on distinctive features for use in an ASR system. He developed a method for representing the speech waveform in terms of a set of abstract linguistic distinctions to derive a set of discriminative features for use in a speech recognizer. He achieved a reduction in the word error rate of 33% on an ASR task by combining the distinctive feature representation with the original waveform representation.

Tolba et al. [26] worked on auditory-based acoustic distinctive features and spectral features for ASR using a multistream paradigm to improve the performance of ASR systems. Their goal was to improve the performance of HMM-based ASR systems by exploiting certain features that characterize speech sounds based on the auditory system and that are based on the Fourier power spectrum. They conducted a series of experiments on speaker-independent continuous speech recognition using a subset of the large readspeech corpus TIMIT. Based on their results they claimed that combining classical MFCCs with auditory-based acoustic features and the main peaks of the spectrum of a speech signal using a multistream paradigm leads to an improvement in the recognition performance. They found that the word error rate decreased by about 4.01%. In addition, they showed that the use of auditory-based acoustic distinctive cues and/or the magnitudes of the spectral peaks improve the performance of the recognition process compared to systems using only MFCCs, their first derivatives, and second derivatives.

Tolba et al. [27] also contributed to another slightly different study. They carried out comparative experiments to evaluate the use of auditory-based acoustic distinctive features and formant cues for ASR using a multistream paradigm. They described an experimental effort to compare the performance of an HMM-based ASR system in which certain speech features were combined with the classical MFCCs using a multistream paradigm. A series of experiments on speaker-independent continuous speech recognition was conducted using a subset of the large readspeech corpus TIMIT. They claimed that the use of either the magnitudes or the frequencies of the speech signals combined with some auditory-based features and MFCCs for ASR using a multistream paradigm leads to an improvement in the recognition performance of ASR systems compared to systems using only MFCCs, and their first and second derivatives. They concluded that combining a perceptual-based front-end with the knowledge gained from measuring the physiological responses to speech stimuli could provide insight into the features used in the auditory system for speech recognition.

Fukuda et al. [28] worked on noise-robust ASR using DPF elements approximated with a logarithmic normal distribution of the HMM. They attempted to replace normal distributions of the DPF elements with logarithmic normal distributions in the HMMs. According to the authors, this is caused by DPF elements showing skew symmetry or positive and negative skewness. They asserted that the proposed HMM with a logarithmic normal distribution yielded a better performance than the HMM with a normal distribution on speakerindependent isolated spoken word recognition tests both with clean speech and speech contaminated by high-level additive noise. Moreover, the combined use of DPF and MFCC significantly improved the word error rate over the baseline HMM system based on the MFCC parameters.

Selouani et al. [29] focused on auditory-based acoustic distinctive features and spectral cues for robust

ASR in low SNR car environments. They proposed a multistream paradigm to improve the performance of ASR systems in the presence of highly interfering car noise. Their results showed that combining classical MFCCs with the main formant frequencies of speech signals using a multistream paradigm leads to an improvement in the recognition performance in noisy car environments for a wide range of SNR values varying between 16 dB and –4 dB. They concluded that the use of auditory-based acoustic distinctive cues could improve the performance of the recognition process in noisy car environments as opposed to using only MFCCs, and their first and second derivatives, for high SNR values, but not for low SNR values.

Stuker et al. [30] addressed the subject of multilingual articulatory features. They addressed articulatory features in the context of monolingual, crosslingual, and multilingual speech recognition. Their results showed that for a variety of languages articulatory features can be reliably recognized within the language and even across languages. They found that pooling the feature detector from multiple languages outperforms monolingual ones. They claimed a relative error rate reduction of 10.7% in a monolingual setup and up to 12.3% in a crosslingual setup.

Fukuda et al. [31] worked on designing multiple DPF extractors for canonicalization using a clustering technique. They showed that the acoustic models of an HMM-based classifier include various types of hidden factors, such as speaker-specific characteristics and acoustic environments. In their work, they described an attempt to design multiple DPF extractors corresponding to unspecific hidden factors, as well as the introduction of a noise suppressor targeted at the canonicalization of a noise factor. Their proposed method isolates the feature extractor design from the HMM classifier design. The Japanese version of the AURORA2 database was used in this work. The authors claimed that the proposed method achieved a significant improvement when combining the canonicalization process with the noise reduction technique based on a 2-stage Wiener filter

Huda et al. [32] worked on the canonicalization of feature parameters for robust speech recognition based on DPF vectors. They introduced a canonicalization method composed of multiple DPF extractors corresponding to each hidden factor canonicalization, and a DPF selector that selects an optimum DPF vector as an input for the HMM-based classifier. They proposed a method to resolve sex factors and speaker variability, and eliminate noise factors by applying the canonicalization based on the DPF extractors and 2-stage Wiener filtering. Having carried out experiments using the Japanese corpus, AURORA-2J, they asserted that the proposed method provides higher word accuracy under clean training and a significant improvement of the word accuracy for low SNR under multicondition training compared to the standard ASR system with MFCC parameters. Moreover, the proposed method requires reduced (two-fifths) Gaussian mixture components and less memory to achieve accurate ASR.

Chen et al. [33] worked on phone set construction based on context-sensitive articulatory attributes for code-switching speech recognition. They presented a novel method for creating a polyglot speech synthesis system via the selection of the speech sample frames of the given speaker in the first language without the need of collecting speech data from a bilingual (or multilingual) speaker. They employed the articulatory features and auditory features in their selection process to achieve a high-quality synthesis output. Moreover, they asserted that the experimental results showed that good performance regarding the similarity and naturalness can be achieved with the proposed method. Finally, Wu et al. [34] tried to create a cross-lingual frame selection method for polyglot speech synthesis. They integrated acoustic features and cross-lingual contextsensitive articulatory features into phone set construction for code-switching ASR by KL-divergence and a hierarchical phone unit clustering algorithm. They claimed that their experimental results show that their method outperforms other traditional phone set construction methods.

4. Conclusion

This work investigated Arabic language research on DPF elements along with some background of this important and specialized language topic. A comparison of Arabic, English, and Japanese phonologies was covered. Various engineering techniques were used aside from the DPF, including MFCC, ANN, HMM, and combinations of these. ANNs are widely used in clustering techniques. Various studies using different techniques were targeted to achieve speech recognition at the highest possible level of performance. Sadly, although Arabic is rich in distinctive features, the quantity and quality of research focusing on DPF construction, analysis, comparison, and adoption in modern digital speech processing and various applications thereof is very limited. We also reviewed a variety of studies that make use of DPFs to realize the highest possible level of performance in speech recognition. These studies employ different engineering techniques to identify the DPFs, including MFCCs, HMMs, ANNs, which are widely used in clustering techniques, and combinations of these.

Acknowledgments

This work is supported by the NPST program by King Saud University Project Number 10-INF1325-02, and Project Number 11-INF1968-02.

References

[1] Y.A. Alotaibi, A.H. Meftah, “Comparative evaluation of two Arabic speech corpora” Natural Language Processing and Knowledge Engineering International Conference, pp. 1–5, 2010.

[2] F. Biadsy, J. Hirschberg, N. Habash, “Spoken Arabic dialect identification using phonotactic modeling”, Proceedings of the European Association for Computational Linguistics, Workshop on Computational Approaches to Semitic Languages, pp. 53–61, 2009.

[3] J. Deller, J. Proakis, J.H. Hansen, Discrete-Time Processing of Speech Signal, London, Macmillan Publishers, 1993.

[4] M. Alkhouli, Alaswaat Alaghawaiyah (Linguistic Phonetics), Daar Alfalah, Jordan, 1990 (in Arabic).

[5] M. Alghamdi, Arabic Phonetics, Al-Toubah Bookshop, Riyadh, 2001 (in Arabic).

[6] K. Kirchhoff, J. Bilmes, S. Das, N. Duta, M. Egan, J. Gang, H. Feng, J. Henderson, L. Daben, M. Noamany, P. Schone, R. Schwartz, D. Vergyri, “Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins summer workshop”, IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 344–347, 2003.

[7] K. Ohata, “Phonological differences between Japanese and English: several potentially problematic areas of pronunciation for Japanese ESL/EFL learners”, Asian English for Specific Purposes Journal, Vol. 6, 2004.

[8] P. Ladefoged, Vowels and Consonants, Second Edition, Oxford, Blackwell Publishing, 2005.

[9] N. Chomsky, M. Halle, The Sound Pattern of English, Massachusetts, MIT Press, 1991.

[10] T. Fukuda, W. Yamamoto, T. Nitta, “Distinctive phonetic feature extraction for robust speech recognition”, IEEE International Conference on Acoustics, Speech, and Signal Processing Vol. 2, pp. II - 25–28, 2003.

[11] T. Nitta, “Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA”, International Conference on Acoustics, Speech, and Signal Processing Vol. 1, pp. 421–424, 1999.

[12] Y.A. El-Imam, “An unrestricted vocabulary Arabic speech synthesis system”, IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. 37, pp. 1829–1845, 1989.

[13] R. Jakobson, G.M. Fant, M. Halle, Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates, Massachusetts, MIT Press, 1963.

[14] E. Eide, “Distinctive features for use in an automatic speech recognition system”, European Conference on Speech Communication and Technology, Vol. 3, pp. 1613–1616, 2001.

[15] M.N Huda, M. Ghulam, J. Horikawa, T. Nitta, "Distinctive phonetic feature (DPF) based phone segmentation using hybrid neural networks", Proceedings of the 8th Annual Conference of the International Speech Communication Association, pp. 94–97, 2007.

[16] M.N. Huda, M. Ghulam, T. Nitta, "DPF based phonetic segmentation using recurrent neural networks", Autumn Meeting of Astronomical Society of Japan, pp. 3–4, 2006.

[17] M. Alghamdi, Arabic Phonetics and Phonology, forthcoming.

[18] S. Selouani, J. Caelen, "Spotting Arabic phonetic features using modular connectionist architectures and a rule-based system", Proceedings of the International ICSC/IFAC Symposium on Neural Computation, pp. 404–411,1998.

[19] S. Selouani, J. Caelen, "Arabic phonetic features recognition using modular connectionist architectures", IEEE 4th Workshop, Interactive Voice Technology for Telecommunications Applications, pp. 155–160, 1998.

[20] S. King, P. Taylor, "Detection of phonological features in continuous speech using neural networks", Computer Speech and Language, Vol. 14, pp. 333–345, 2000.

[21] B. Launay, O. Siohan, A. Surendran, C. Leet, "Towards knowledge-based features for HMM based large vocabulary automatic speech recognition", IEEE International Conference on Acoustics, Speech, and Signal Processing pp. I-817–I-820, Vol. 1, 2002.

[22] T. Fukuda, T. Nitta, "Canonicalization of feature parameters for automatic speech recognition", International Conference on Spoken Language Processing Vol. 4, pp. 2537–2540, 2004.

[23] M.N Huda, M. Ghulam, K. Katsurada, Y. Iribe, T. Nitta, "Distinctive phonetic feature (DPF) based phone segmentation using 2-stage multilayer neural networks", The Research Institute of Signal Processing, International Workshop on Nonlinear Circuits and Signal Processing pp. 325–328, 2007.

[24] M.N Huda, K. Katsurada, T. Nitta, "Phoneme recognition based on hybrid neural networks with inhibition/enhancement of distinctive phonetic feature (DPF) trajectories", Proceedings of the 9th Annual Conference of the International Speech Communication Association, pp. 1529–1532, 2008.

[25] D. Yu, S.M. Siniscalchi, L. Deng, CH. Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition", International Conference on Acoustics, Speech, and Signal Processing pp. 4169–4172, 2012.

[26] H. Tolba, S. Selouani, D. O'Shaughnessy, "Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm", International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. I-837–I-840, 2002.

[27] H. Tolba, S.A. Selouani, D. O'Shaughnessy, "Comparative experiments to evaluate the use of auditory-based acoustic distinctive features and formant cues for automatic speech recognition using a multi-stream paradigm", Proceeding of the 7th International Conference on Spoken Language Processing, pp. 113–2116, 2002.

[28] T. Fukuda, T. Nitta, "Noise-robust ASR by using distinctive phonetic features approximated with logarithmic normal distribution of HMM", European Conference on Speech Communication and Technology, Vol. 3, pp. 2185–2188, 2003.

[29] S. Selouani, H. Tolba, D. O'Shaughnessy, "Auditory-based acoustic distinctive features and spectral cues for robust automatic speech recognition in low-SNR car environments", Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Companion Volume of the Proceedings of HLT-NAACL, Vol. 2, pp. 91–93, 2003.

[30] S. Stüker, T. Schultz, F. Metze, A. Waibel, "Multilingual articulatory features", IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings, Vol. 1, pp. I-144–I-147, 2003.

[31] T. Fukuda, T. Nitta, "Designing multiple distinctive phonetic feature extractors for canonicalization by using clustering technique", European Conference on Speech Communication and Technology, pp. 3141–3144, 2005.

[32] M.N Huda, M. Ghulam, T. Fukuda, K. Katsurada, T. Nitta, "Canonicalization of feature parameters for robust speech recognition based on distinctive phonetic feature (DPF) vectors", The Institute of Electronics, Information and Communication Engineers Journal, Vol. E91–D, pp. 488–498, 2008.

[33] CP. Chen, YC. Huang, CH. Wu, KD Lee, “Cross-lingual frame selection method for polyglot speech synthesis”, International Conference on Acoustics, Speech, and Signal Processing pp. 4521–4524, 2012.

[34] CH. Wu, HP. Shen, YT. Yang, “Phone set construction based on context-sensitive articulatory attributes for code-switching speech recognition”, International Conference on Acoustics, Speech, and Signal Processing pp. 4865–4868, 2012.