日中韓辭典研究所
The CJK Dictionary Institute

Enhancing Arabic Speech Technology

with comprehensive Arabic training lexicon

by Jack Halpern
jack@cjki.org

Recent advances in deep learning and neural network technology have dramatically improved speech technology [5], but these improvements are not uniform and some languages, due to their complexity, unfortunately lag behind. Arabic is one such language, and although universally regarded as one of the world's major languages, the quality of Arabic speech technology is still significantly behind that of the other major languages such as Chinese, Spanish, and English. This report analyzes errors in the Arabic TTS systems for Google, Apple, and Microsoft, and explains how **ArabLEX**, CJKI's **Comprehensive Arabic Full Form Lexicon** [1, 2], the most comprehensive Arabic computational lexicon available, can serve as training lexicon to address these shortcomings and enhance the quality of both text-to-speech (TTS) and automatic speech recognition (ASR).

1. TTS accuracy

The extreme orthographic ambiguity of Arabic (see Appendix 2 for linguistic details) has led to very high error rates in TTS systems, even those offered by major companies such as Google, Apple, and Microsoft. We tested several of these TTS systems to determine the scope of the problem and present some of the results in Appendix 1. We discovered that it is not unusual for over 50% of the words in a sentence in these systems to be mispronounced, and that there is a trend for cliticized words to be incorrectly pronounced. For example, the cliticized word وَلِلْكَاتِبِينَ, correctly pronounced *walilkạtibīna,* is mispronounced as *walilkātibáyna.*

Another issue is prosody (stress and intonation) and vowel neutralization (e.g. نا *nạ* is written as a long vowel in أنا but is shortened in actual pronunciation to *na*). This means that the ي in both يا and يد are pronounced long and short respectively, as it seems that they should be, but in fact the long vowel in the former should be neutralized and both are

pronounced *yạ* (the *ạ* indicates a shortened long vowel). Incorrect prosody means for example pronouncing مدرسة 'school' as *mádrasatu* instead of the correct *madrásatu.* This causes the Arabic to sound unnatural, like pronouncing *nation* with the stress on the last syllable. This complex issue is described more fully in my paper on Arabic stress [3].

2. ASR accuracy

ArabLEX includes features specifically designed to support automatic speech recognition (ASR). For speech synthesis (TTS), it is necessary to *generate* only one accurate pronunciation. For example, كاتبون 'writers' in standard Arabic is pronounced *kạtibūna,* but for ASR it is also necessary to *recognize* the less formal variant pronunciation *kạtibūn*. Similarly, the standard pronunciation of أكتب 'I write' is *ʾáktubu,* but the final vowel is often omitted and it is pronounced *ʾáktub.*

The above alternatives are on a *phonemic* level. That is, the phoneme /na/ is being replaced by the phoneme /n/ as a result of vowel omission. There are also variations on the *phonetic* level; that is, certain phonemes have regional allophones. For example, ج in such words as جمل *jamal* is pronounced [gɛ̈mɛl] in Egypt, [d̲ʒɛ̈mɛ̈l] in the Gulf region, and [ʒɛ̈mɛ̈l] in the Levant. It is important to note that this does not refer to the local dialects in those regions, but to a *regional varieties* of Modern Standard Arabic (MSA) in those regions.

Thus, *ArabLEX* not only represents ج in the standard IPA [d̲ʒ] for TTS, it also lists the regional [ʒ] and [g] for ASR training. The goal is to enable the recognition of these allophones, but not to generate them.

3. Comprehensive pronunciation dictionary

One of the key components for training speech technology systems is the *pronunciation dictionary.* A major feature of *ArabLEX* is that it can serve as an extremely comprehensive pronunciation dictionary as well as training lexicon.

ArabLEX not only maps all unvocalized forms to their vocalized counterparts and to their

lemmas, but also provides precise **phonemic transcriptions** (CARS system) [4] and **phonetic transcriptions** (IPA) that includes precise word stress and vowel neutralization. For example, in *wɛ̈likɛ̈ːˈtibikumɛ(ˑ)*, the stressed syllable is indicated by (ˈ) (U+0C28), while (ˑ) (U+02D1) indicates that the final ɛ is neutralized vowel of optional half length. This helps developers significantly enhance the quality of Arabic TTS, and can be used in training ASR systems to achieve higher recognition rates.

4. Conclusion

To summarize, *ArabLEX* can bring the following benefits to Arabic speech technology:

- Approximately 600 million entries, including millions of proper nouns.
- *All combinations* of proclitics, enclitics, and affixes.
- Tens of millions of orthographic variants.
- Exhaustive list of alternative pronunciations of identical unvocalized strings to enable orthographical disambiguation (e.g. six alternatives for كاتباتك).
- Highly accurate phonemic transcriptions for all wordforms (CARS).
- Phonetic transcriptions (IPA) indicate the correct allophonic variants in context as well as regional variants.

Speech synthesis, speech recognition and prosody in current Arabic speech technology are, on the whole, inaccurate, unnatural and often unpleasant to the ear. Although Arabic is one of the most common languages in the world, Arabic speech technology has lagged behind the other major world languages. This does a disservice to Arabic-speaking users, not to mention the many millions of people around the world who are learning Arabic and rely on speech tools to further their learning and communicative ability.

References

[1] Halpern, Jack. Arablex: Comprehensive Arabic Full Form Lexicon
[2] Halpern, Jack. Very Large-Scale Lexical Resources to Enhance Chinese and Japanese Machine Translation
[3] Halpern, Jack. Word Stress and Vowel Neutralization in Modern Standard Arabic
[4] Halpern, Jack. CJKI Arabic Romanization System
[5] Wajdan Algihab, Imam Muhammad bin Saud Islamic University. Arabic Speech Recognition with Deep Learning: A Review

Appendix 1: TTS TEST RESULTS

Below are the results of CJKI's tests comparing the TTS systems of Google, Apple (iOS) and Microsoft (Bing). The **Unvocalized** field is the original Arabic text, the **Vocalized** field indicates the correct pronunciation, and the **CJKI** field shows the pronunciation in CARS phonemic transcription [4] contained in *ArabLEX*. The CARS transcriptions in these **Google, iOS** and **Bing** columns indicate the actual pronunciation by the three TTS engines. Mispronunciations are indicated in red, and the error rate is given in the column headers.

Tables 1 and 3 are based on text **composed** for this survey, while tables 2 and 4 use a sentence **extracted** from the web. (It is noteworthy that the error rate for the composed text is actually much lower than for the extracted text.) Tables 1 and 3 compare the results on a word-by-word basis, whereas tables 2 and 4 compare them on a sentence-by-sentence basis, showing the context.

Table 1: Mispronounced Words in Composed Text

Unvocalized	Vocalized	Google (13%)	iOS (31%)	Bing (25%)	CJKI
عدد	عَدَّدَ	ɛádadu	ɛádada	ɛádada	ɛáddada
الكاتب	ٱلْكَاتِبُ	lkātibu	lkātibi	lkātibu	lkātibu
ما	مَا	mạ	mạ	mạ	mạ
قال	قَالَ	qāla	qāla	qāla	qāla
إن	إِنَّ	ʾínna	ʾínna	ʾínna	ʾínna
هؤلاء	هٰؤُلَاءِ	hạʾulāʾi	hạʾulāʾi	hạʾulāʾi	hạʾulāʾi
الحكام	ٱلْحُكَّامَ	lḥukkāmi	lḥukkāmi	lḥukkāmi	lḥukkāma
يفعلونه	يَفْعَلُونَهُ	yafɛalūnahu	yafɛalūnahu	yafɛalūnahu	yafɛalūnahu
في	فِي	fi	fi	fi	fi
الخارج	ٱلْخَارِجِ	lkhāriji	lkhārija	lkhāriji	lkhāriji
مثل	مِثْلَ	míthli	míthli	míthli	míthla

الهجمات	ٱلْهَجَمَاتِ	lhajamāti	lhajamāti	lhajamāti	lhajamāti
الإلكترونية	ٱلْإِلِكْتُرُونِيَّةِ	lʾilikturu̱níyyati	lʾilikturu̱níyyati	lʾilikturu̱níyyati	lʾilikturu̱níyyati
ومطاردة	وَمُطَارَدَةِ	wamuṭārádati	wamuṭārídati	wamuṭārídati	wamuṭārádati
المعارضين	ٱلْمُعَارِضِينَ	lmuɛa̱riḍīna	lmuɛa̱riḍīna	lmuɛa̱riḍīna	lmuɛa̱riḍīna
اللاجئين	ٱللَّاجِئِينَ	llaji’īna	llaji’īna	llaji’īna	llaji’īna
في	فِي	fi̱	fi̱	fi̱	fi̱
العواصم	ٱلْعَوَاصِمِ	lɛawāṣimi	lɛawāṣimi	lɛawāṣimi	lɛawāṣimi
الغربية	ٱلْغَرْبِيَّةِ	lgharbíyyati	lgharbíyyati	lgharbíyyati	lgharbíyyati
وللكاتبين	وَلِلْكَاتِبِينَ	walilka̱tibīna	walilkātibáyna	walilkātibáyna	walilka̱tibīna
من	مِنَ	mína	mína	mína	mína
الصحفيين	ٱلصَّحَفِيِّينَ	ṣṣaḥafi̱yīna	ṣṣaḥafi̱yīna	ṣṣaḥafi̱yīna	ṣṣaḥafi̱yīna
العرب	ٱلْعَرَبِ	lɛárabi	lɛárabi	lɛárabi	lɛárabi
صرح	صَرَّحَ	ṣárraḥa	ṣáraḥa	ṣáraḥa	ṣárraḥa
بأن	بِأَنَّ	biʾánna	biʾánna	biʾánna	biʾánna
عليهم	عَلَيْهِمْ	ɛaláyhim	ɛaláyhim	ɛaláyhim	ɛaláyhim
أن	أَنْ	ʾan	ʾan	ʾan	ʾan
يكتبوا	يَكْتُبُوا	yaktúbuwu	yaktúbuwu	yaktúbuwu	yaktúbuwu
ما	مَا	ma̱	ma̱	ma̱	ma̱
تمليه	تُمْلِيهِ	tumalīhi	tamlīhi	tamlīhi	tumlīhi
عليهم	عَلَيْهِمْ	ɛalayhim	ɛalayhim	ɛalayhim	ɛalayhim
ضمائرهم	ضَمَائِرُهُمْ	ḍamāʾíruhum	ḍamāʾírihim	ḍamāʾírihim	ḍamāʾíruhum

Table 2: Mispronounced Words in Extracted Text

Unvocalized	Vocalized	Google (80%)	iOS (90%)	Bing (70%)	CJKI
الاخوات	ٱلْأَخَوَاتُ	ʾalikhwātu	ʾalʾakhawāti	ʾalʾakhawātu	ʾalʾakhawātu
المتزوجات	ٱلْمُتَزَوِّجَاتُ	lmutazawwijātu	lmutazawwijāti	lmutazawwijāti	lmutazawwijātu
اللاتي	ٱللَّاتِي	lṭi	llaṭi	llāṭi	llāṭi
رزقن	رُزِقْنَ	rízqin	rúzqin	rúzqin	ruzíqna
بابناء	بِأَبْنَاءَ	bạbināʾun	bibnāʾi	bibnāʾi	biʾabnāʾa
فليكتبن	فَلْيَكْتُبْنَ	falayiktíbna	falktíbna	falktíbna	falyaktúbna
اسمائهم	أَسْمَائِهِمْ	ʾismāʾahum	smāʾihim	smāʾihim	ʾasmāʾahum
وسبب	وَسَبَبَ	wasábaba	wasábaba	wasábaba	wasábaba
التسميه	ٱلتَّسْمِيَةِ	lttasammīhu	lttasammīhi	lttasammīhi	ttasmíyati
رجاءا	رَجَاءًا	rajjāʾan	rajāʾ	rajāʾ	rajāʾan

Table 3: Mispronounced Sentences in Composed Text

TTS	Sentence	Error %
Unvocalized	عدد الكاتب ما قال إن هؤلاء الحكام يفعلونه في الخارج مثل الهجمات الإلكترونية ومطاردة المعارضين اللاجئين في العواصم الغربية. وللكاتبين من الصحفيين العرب صرح بأن عليهم أن يكتبوا ما تمليه عليهم ضمائرهم	-
Vocalized	عَدَّدَ ٱلْكَاتِبُ مَا قَالَ إِنَّ هَٰؤُلَاءِ ٱلْحُكَّامَ يَفْعَلُونَهُ فِي ٱلْخَارِجِ مِثْلَ ٱلْهَجَمَاتِ ٱلْإِلِكْتُرُونِيَّةِ وَمُطَارَدَةَ ٱلْمُعَارِضِينَ ٱللَّاجِئِينَ فِي ٱلْعَوَاصِمِ ٱلْغَرْبِيَّةِ. وَلِلْكَاتِبِينَ مِنَ ٱلصَّحَفِيِّينَ ٱلْعَرَبِ صَرَّحَ بِأَنَّ عَلَيْهِمْ أَنْ يَكْتُبُوا مَا تُمْلِيهِ عَلَيْهِمْ ضَمَائِرُهُمْ	-
CJKI	ɛáddada lkātibu mạ qāla ʾínna hạ̄ʾulāʾi lḥukkāma yafɛalūnahu fi lkhāriji míthla lhajamāti lʾilikturụníyyati wamuṭārádati lmuɛạriḍīna llạjiʾīna fi lɛawāṣimi lgharbíyyati. walilkạtibīna mína ṣṣaḥafiyīna lɛárabi ṣárraḥa biʾánna	-

	ɛaláyhim ʾan yaktúbuwu mạ tumlīhi ɛalayhim ḍamāʾíruhum	
Google	ɛádadu lkātibu mạ qāla ʾínna hạʾulāʾi lḥukkāmi yafɛalūnahu fi̱ lkhāriji míthli lhajamāti l’ilikturụníyyati wamuṭārádati lmuɛạriḍīna llạji῾īna fi̱ lɛawāṣimi lgharbíyyati. walilkạtibīna mína ṣṣaḥafi̱yīna lɛárabi ṣárraḥa biʾánna ɛaláyhim ʾan yaktúbuwu mạ tumallīhi ɛalayhim ḍamāʾíruhum	13%
iOS	ɛádada lkātibi mạ qāla ʾínna hạʾulāʾi lḥukkāmi yafɛalūnahu fi̱ lkhārija míthli lhajamāti l’ilikturụníyyati wamuṭārídati lmuɛạriḍīna llạji῾īna fi̱ lɛawāṣimi lgharbíyyati. walilkātibáyna mína ṣṣaḥafi̱yīna lɛárabi ṣáraḥa biʾánna ɛaláyhim ʾan yaktúbuwu mạ tamlīhi ɛalayhim ḍamāʾírihim	31%
Bing	ɛádada lkātibu mạ qāla ʾínna hạʾulāʾi lḥukkāmi afɛalūnahu fi̱ lkhāriji míthli lhajamāti l’ilikturụníyyati wamuṭārídati lmuɛạriḍīna llạji῾īna fi̱ lɛawāṣimi lgharbíyyati. walilkātibáyna mína ṣṣaḥafi̱yīna lɛárabi ṣáraḥa biʾánna ɛaláyhim ʾan yaktúbuwu mạ tamlīhi ɛalayhim ḍamāʾírihim	25%

Table 4: Mispronounced Sentences in Extracted Text

TTS	Sentence	Error %
Unvocalized	الاخوات المتزوجات اللاتي رزقن بابناء فليكتبن اسمائهم وسبب التسميه رجاءا	-
Vocalized	اَلْأَخَوَاتُ ٱلْمُتَزَوِّجَاتُ ٱللَّاتِي رُزِقْنَ بِأَبْنَاءَ فَلْيَكْتُبْنَ أَسْمَائَهُمْ وَسَبَبَ ٱلتَّسْمِيَةِ رَجَاءا	-
CJKI	ʾalʾakhawātu lmutazawwijātu llāti̱ ruzíqna biʾabnāʾa falyaktúbna ʾasmāʾahum wasábaba ttasmíyati rajāʾan	-
Google	ʾalikhwātu lmutazawwijātu lti̱ rízqin bạbināʾun falayiktíbna ʾismāʾahum wasábaba lttasammīhu rajjāʾan	80%
iOS	ʾalʾakhawāti lmutazawwijāti llati̱ rúzqin bibnāʾi falktíbna smāʾihim wasábaba lttasammīhi rajāʾ	90%
Bing	ʾalʾakhawātu lmutazawwijāti llāti̱ rúzqin bibnāʾi falktíbna smāʾihim wasábaba lttasammīhi rajāʾ	70%

Appendix 2: Orthographical ambiguity

One reason that Arabic speech technology lags behind is that the Arabic script is highly ambiguous. Words are often written as a string of consonants with no indication of vowels. For example, كاتب can represent as many as *seven* pronunciations: *kāatib*, *kātibun*, *kātibin*, *kātaba*, *kātibi*, *kātiba* and *kātibu*. Many other characteristics of the Arabic script contribute to a high level of orthographic ambiguity, as described in Halpern's paper on Arabic named entities [4].

The morphological complexity of such cliticized forms as ولكاتباتهما *walikạtibātíhimạ*, and the absence of vowel diacritics, makes Arabic TTS especially challenging. That is, determining the morphological composition of such forms, and the correct vowels for such consonants as ت in ولكاتباتهما often requires morphological, semantic and contextual analysis which tax the capabilities of state-of-the-art speech technology.