THE ROLE OF PHONETICS AND PHONETIC DATABASES IN JAPANESE SPEECH TECHNOLOGY

Jack Halpern (春遍雀來)

The CJK Dictionary Institute (日中韓辭典研究所)
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001, Japan

ABSTRACT

This paper summarizes the complex allophonic variations that need to be considered in developing Japanese speech technology applications, with special emphasis on natural speech synthesis, and describes a large-scale phonetic database that provides IPA transcriptions with accent codes and including hundreds of thousands of allophones. The common notion that the kana syllabary correctly represents the Japanese phoneme stock is only partially true. A truly phonemic orthography requires unambiguous grapheme-to-phoneme mappings. Yet speech technology developers often use kana pronunciation dictionaries because of easy availability and low cost. Even more challenging is to accurately generate the allophonic variants and accent patterns for each phoneme and lexeme, which has been done in the project reported in this paper.

Index Terms**—** speech synthesis (TTS), speech recognition (ASR), Japanese allophones, IPA, pitch accent

1. INTRODUCTION

1.1. Overview

The popular notion that the kana syllabary is "phonetic" (*phonemic* would be the right term), that is, that it correctly represents the Japanese phoneme stock, is only partially true. In a truly phonemic orthography, there is an unambiguous one-to-one grapheme-to-phoneme mapping, but the kana syllabary is not truly phonemic, even if modified to eliminate such ambiguities as /u/ representing [ɯ] or [o]. Yet Japanese speech technology developers often use kana pronunciation dictionaries because of easy availability and low cost. This is the first hurdle that needs to be overcome in developing Japanese speech synthesis systems. The central difficulty is to accurately convert phonemes into phones by generating the correct allophonic variants and accent patterns for each phonetic environment.

This paper summarizes the complex allophonic variation patterns in Japanese, such as vowel devoicing, nasalization and affricate spirantization, that need to be considered in implementing such applications as natural speech synthesis, speech recognition for generating allophonic variants for pronunciation dictionaries, and phonetic transcription systems (such as kana-to-IPA). Though this paper does not cover any issue in-depth, it does clarify the difficult challenges, from a phonetic point of view, that must be tackled in achieving natural speech synthesis.

The paper also describes a 150,000-entry **Japanese Phonetic Database** (JPD) that provides, for the first time, both IPA transcriptions and accent codes. It is designed to contribute to the advancement of Japanese speech technology by providing IPA transcriptions that accurately indicate how Japanese words and personal names are pronounced in actual speech, and covers hundreds of thousands of allophonic variation patterns [1].

1.2 Practical Applications

The practical applications of JPD include:

1. Text-to-speech (TTS) systems (speech synthesis) based on a database of phonetic transcriptions.
2. Automatic speech recognition (ASR) systems (generating allophones for pronunciation dictionaries).
3. Pedagogical research to aid in the acquisition of Japanese as a foreign language.
4. Research and development of Japanese speech technology in general.
5. Phonetic transcription systems, such as kana-to-IPA.

An important motivation for the dependence of speech technology on statistical techniques has been the lack of phonetic databases with phonetic transcriptions and/or accent codes for each Japanese lexeme. Below is a brief discussion of how JPD can be applied to TTS and ASR.

Contemporary approaches to Japanese TTS use statistical methods such as hidden Markov models (HMMs) representing triphones to model the allophonic variation that occurs in different phonetic contexts [2]. In such models each component phone is based on a phone-set derived directly from the kana syllabary with rule-based adjustments to correct the most common grapheme-to-phone conversion errors. Using the JPD for Japanese TTS can facilitate the development of monophone systems exploiting JPD's detailed IPA phone-set. For this to be effective, it is expected that a statistical framework, such as HMMs, will also be necessary.

Japanese ASR employs a similar method to that described for TTS. It is unlikely that the JPD will be used directly in the development of acoustic models since this would require the vast amounts of existing audio data to be retranscribed using the same IPA phone-set. This would obviously be prohibitively time-consuming and expensive. However, the JPD could be used directly to expand existing kana-based pronunciation dictionaries with a wider range of allophones. Similarly, the JPD could be used to produce more accurate kana-to-phone conversion tools to produce more accurate phonetic transcriptions for training acoustic models. Both the increased number of allophones and the improved automatic phonetic transcriptions can be expected to be beneficial for acoustic model training to give both improved models as well as greater accuracy during recognition.

1.3. Conventions

The orthographic conventions used in this paper are illustrated by 新聞 *shimbun* 'newspaper':

- Slashes are used for phonemic transcription, e.g. /siNbuN/.
- Katakana is used for kana orthography, e.g. シンブン.
- Square brackets are used for phonetic transcription in IPA. e.g. [ɕimbɯɴ].
- Italics are used for Hepburn romanization, e.g. *shimbun*.
- Single quotes are used for English equivalents of Japanese words, e.g. 'newspaper'.

A word of caution regarding transcription conventions. The five standard Japanese vowels are transcribed in IPA using broad transcription, as shown in the table below. This is not a precise representation of the actual pronunciation. For example, both エ /e/ and オ /o/ are lower than the cardinal vowels [e] and [o], so that [e̞] and [o̞] are closer. In particular, the realization of ウ /u/ is variable, as shown in the **Narrow** column in Table 1, and cannot be rendered precisely by any IPA symbol. It is a somewhat fronted back vowel pronounced with compressed lips, and is different from the close back unrounded vowel represented by [ɯ]. Nevertheless, /u/ is traditionally transcribed by [ɯ], and we will follow that convention.

Table 1: Japanese Vowels

Kana	Phonemic	Broad	Narrow
ア	a	a	ä
イ	i	i	i
ウ	u	ɯ	ɯ̹̣̈ ~ ü
エ	e	e	e̞
オ	o	o	o̞

Another issue is how to transcribe fricatives/affricates such as シ and チ and ジ. Traditionally, perhaps through the influence of English, these are often transcribed as [ʃi], [tʃi] and [dʒi], but we transcribe them as [ɕ], [tɕi] and [dʑi] because strictly speaking they are pronounced slightly further back than the corresponding English ones; that is, they are alveolopalatal rather than palatoalveolar (an even more precise rendering could be [cɕi] for チ and [ɟʑi] for ジ).

2. IS KANA A PHONEMIC ORTHOGRAPHY?

Though on the whole the kana syllabary is fairly phonemic, that is, each kana symbol represents one phoneme (such as ア = /a/) or a specific sequence of two phonemes (such as カ = /k/ + /a/), kana can be ambiguous, such as ウ representing either /o/ or /u/. For example, ウ in トウ in 塔 'tower' represents the phoneme /o/ (phonetically [o]), elongating the previous /o/, while in トウ as the reading of 問う 'ask' represents /u/ (phonetically [ɯ]), two distinct phonemes.

The kana syllabary is not a true phonemic orthography because of various one-to-many ambiguities, such as:

1. Variation in long vowel representation, e.g. both ウ or オ are used for long /o/, as in トオリ (通り 'road') and トウリ (党利 'party interests') and エ or イ are used for long /e/, as in ネエサン /neesan/ (姉さん 'older sister') and ケイサン /keesan/ (計算 'calculation').
2. Similarly, イ can represent either /i/ or /e/ in ケイ、メイ etc. For example, メイン 'main' can be /mein/ or /meen/.
3. Historical kana ambiguity, such as ハ representing both /ha/ or /wa/. For example, the topic marker /wa/ is represented by ハ, as in ワタシハ (私は 'I'), which is exactly the same sound as ワ /wa/.
4. Some kana, such as ジ/ヂ /zi/ and ヅ/ズ /zu/, represent exactly the same phoneme. For example, /zi/ (phonetically [dʑi]) is normally written ジ as in ジブン (自分 'self') but as ヂ in チヂム (縮む 'shrink').
5. Allophonic alternation, such as the phoneme /t/ representing [tɕ] before /i/ or [ts] before /u/, e.g. チ /ti/ is [tɕi], /tu/ is [tsu] while タ /ta/ is [ta].

3. ALLOPHONIC VARIATION

Grapheme-to-phoneme ambiguity is only half the battle – the easy half. The real challenge is to convert the phonemes into phones. Since Japanese, like all other languages, has allophonic variation that depends on the phonetic environment, there is often no phoneme-to-phone correspondence on a one-to-one basis. Accurate phoneme-to-phone transformation is a prerequisite to natural speech synthesis.

The environments that lead to allophonic variation in Japanese are complex. Though Japanese speech technology developers often use kana pronunciation dictionaries, kana (even if modified to eliminate one-to-many ambiguities such as by adding diacritics to indicate if ウ represents [ɯ] or [o]), cannot accurately represent the phones as they are actually realized in various environments. Following is a brief description of the main allophonic and other phonetic changes, such as devoicing of vowels and nasalization of moraic /N/, which occur mostly unconsciously in the natural speech of native speakers. These have a marked effect on the naturalness of synthesized speech, and need to be considered in the development of speech synthesis technology.

3.1. Vowel Devoicing

A salient feature of the standard Tokyo dialect is that the unaccented high vowels /i/ and /u/ tend to be devoiced between voiceless consonants and other environments [3]. For example, /su/ in /ainosuke/ (愛之助 proper noun) is realized as a devoiced [sɯ̥], rather than the normal [sɯ]. Optionally, devoicing also occurs word finally after voiceless consonants, so that 続く/tuzuku/ (‘to continue’) could be either [tsɯ̥zɯkɯ̥] or [tsɯ̥zɯkɯ]. Devoicing may also occur for /o/ and even /a/, but these are of lesser importance. Sometimes, what is traditionally considered devoicing in Japanese phonetics is actually a total lack of vowels. For example, the /su/ in /nan desu ka/ is realized as vowelless [s], rather than the devoiced [sɯ̥].

3.2. Nasalization of /g/

/g/ is pronounced as [g] word initially, but is often nasalized word-internally and realized as [ŋ] (for some speakers as [ŋg]), a phonetically distinct allophone of /g/ which is normally realized as the voiced velar stop [g]. For example, in /kage/ (影 ‘shadow’), /ge/ represents a nasalized allophone of /ge/ realized as [ŋe], so it is pronounced [kaŋe]. Nasalized /g/ is gradually falling into disuse among the younger generations. For some speakers, especially in fast speech, intervocalic /g/ in certain environments is realized as a voiced velar fricative [ɣ], so that /kage/ becomes [kaɣe].

Table 2: Nasalization Variants

Orthographic	Kana	Phonetic	Remarks
影	カゲ	[kage]	unnasalized by most speakers
影	カゲ	[kaŋe]	nasalized intervocalically
影	カゲ	[kaɣe]	fricativized in fast speech
画像	ガゾウ	[gazoː]	never nasalized word initially

3.3. Nasal Assimilation

The phoneme /N/ (ン), a moraic nasal, is realized as six different allophones governed by complex rules of nasal assimilation involving coarticulation. Phonetically /N/ behaves as the most complex phoneme in the Japanese phoneme stock. For natural speech synthesis it is important to generate the correct allophones of /N/.

Even linguistically naive native speakers, for whom allophonic variation is mostly unconscious, notice that /N/ is pronounced as [m] in certain environments, such as in /siNbuN/ (新聞シンブン), realized as [ɕimbɯɴ]. But other /N/ allophonic rules are subtle and mostly go unnoticed by native speakers. The most important rules (somewhat informally) are that /N/ is realized as a bilabial nasal [m] when followed by a bilabial [m], [b] or [p], as a velar nasal [ŋ] when followed by the velar stops /k/ and /g/, as [n] before /t/, /d/, /n/ and /r/, as a nasalized vowel of various qualities before vowels, semivowels and some consonants, and as an uvular nasal [ɴ] word finally. The examples below illustrate the six different realizations of /N/.

Table 3: Examples of Nasal Assimilation

Orthographic	Kana	Phonetic
自分	ジブン	dʑibɯɴ
純子	ジュンコ	dʑɯŋko
慎一	シンイチ	ɕiĩtɕi̥
新聞	シンブン	ɕimbɯɴ
運動	ウンドウ	ɯndoː
蒟蒻	コンニャク	koɲɲakɯ
本	ホン	hoɴ, hõ

3.4. Spirantization of Affricates

A subtle feature of Japanese allophonic variation is the spirantization of certain affricates occurring word-internally.

For example, /zi/ is realized as an alveolopalatal affricate [dʑi] word-initially, as in [dʑibɯɴ] (自分 'self'), but intervocalically as the alveolopalatal fricative [ʑi] (or sometimes [dʑi]), as in [haʑi] (恥 'shame'). Similarly, /zu/ is realized as [dzu] word initially but as [zu] in other positions. A related phenomenon is a tendency, especially among young Tokyo females, to pronounce /s/ as [si], an alveolar fricative, rather than the normal alveolopalatal fricative [ɕi].

Table 4: Spirantized Affricates

Orthographic	Phonemic	Phonetic
自分	/zibun/	[dʑibɯɴ (ɟʑibɯɴ)]
恥	/hazi/	[haʑi (haᵈʑi)]
地震	/zisiɴ /	[dʑiiɴ (ɟʑiçiɴ)]
塩	/sio/	[ɕio (sio)]

Spirantization is essentially a kind of weakening (lenition) of affricates in most non-initial positions, the degree of which depends on the speaker, so that fricatives/affricates alternation can be said to be in free variation.

3.5. Sequential Voicing

Sequential voicing is a common phenomenon, governed by complex rules, such as the frequent voicing of the initial consonant of the second element of a compound, e.g. the [su] in 寿司 'sushi' is pronounced [sɯɕi] in isolation but [zɯ] in the compound いなり寿司 'inarizushi', pronounced [inaɾizɯɕi].

Table 5: Sequential Voicing

Orthographic	Voiceless	Voiced
いなり寿司	[inaɾi] + [sɯɕi]	[inaɾizɯɕi]
物語	[mono] + [kataɾi]	[monogataɾi]
二人連れ	[futaɾi] + [tsuɾe]	[futaɾizuɾe]
棒立ち	[boː] ＋[tatɕi]	[boːdatɕi]
花火	[hana] ＋[hi]	[hanabi]

The last item in the table shows a change from the glottal fricative [h] the bilabial stop [b], phonetically unrelated to voicing but a common sequential voicing phenomenon. Attempting to predict sequential voicing by rules is futile as it is often unpredictable. The only safe way is to store such voiced lexemes in a hardcoded database.

3.6. Palatilization

Certain consonants followed by /j/ are palatalized. This is represented in kana by small ャ,ュ and ョ, as in ギャ [gʲa], ギュ [gʲɯ] and ギョ [gʲo]. Certain consonants followed by /i/, especially /n/, are also palatalized, so that /niQpon/ (日本 'Japan') is pronounced [nʲip̚poɴ]. Light palatilization may also occur in such phonemes as /mi/ and /ki/, as shown below.

Table 6: Palatalized Consonants

Orthographic	Phonemic	Phonetic
客	/kyaku/	[kʲaku]
日本	/niQpon/	[nʲip̚poɴ]
民	/tami/	[tamʲi]
滝	/taki/	[takʲi]

3.7. Consonant Gemination

Geminated consonants in Japanese consists of two identical consonants (moraic obstruents) interrupted by a pause, with each consonant belonging to a different mora. This is represented in kana by small ッ. According to the mainstream interpretation of Japanese phonology, this is phonemicized as an archiphoneme represented by /Q/. For example, /tatta/ (立った 'stood') becomes /taQta/.

/Q/ represents a variety of phonetically distinct sounds depending on the following consonant. In cases like /haQsai/ (八歳 'eight years old'), it doubles the consonant and is realized as [hassai], but in cases like /taQta/, the quality of the first [t] is different because it is unreleased, whereas the second [t] is a normal stop. Strictly speaking, /taQta/ is thus realized as [tat̚ta], rather than [tatta]. Moreover, if the geminated consonant is an affricate, only the plosive portion of the affricate is repeated, so that /haQtjuu/ (発注 'ordering goods') is realized as [hat̚tɕɯː], not as [hatɕtɕɯː].

Table 7: Geminated Consonants

Orthographic	Phonemic	Phonetic
日本	/niQpon/	[nʲip̚poɴ]
八歳	/haQsai/	[hassai]
発注	/haQtuu/	[hat̚tɕɯː]
発車	/haQsja/	[haɕɕa]

3.8. Vowel Glottalization

Japanese vowels are sometimes preceded or followed by a glottal stop, often in short words standing alone, and for emphasis. In some cases, the word final glottal stop is clearly audible and is represented orthographically by a small ッ, as in アッ /aQ/ (アッ 'Oh!'), pronounced [aʔ].

Table 8: Glottalized Vowels

Orthographic	Phonemic	Phonetic
アッ	/aQ/	[aʔ̚]
鵜	/u/	[ʔɯʔ]
サッ	[saQ]	[saʔ]

4. PITCH ACCENT

The accentual system of Japanese is a mora-based pitch accent, which is distinct from typical tone languages like Chinese. In Chinese, the tone for each syllable must be specified, whereas in Japanese it is only necessary to specify the accented mora, from which the pitch pattern of the entire word can be determined by phonological rules.

For example, in /anata/ (あなた 'you'), the second mora /na/ is high pitched. All morae following the accented one are lowered. In addition, there is a rule that the first mora is always lowered, unless it is the accented one. This means that /anata/ gets a pitch pattern of LHL (low-high-low), pronounced [anáta]. If no accent is specified, the word is considered accentless, as in /katati/ (形 'shape'), but lowering the pitch of the first mora results in a pitch pattern of LHH.

Below are examples of Japanese accent patterns. The "L" or "H" in parentheses indicates the pitch of particles immediately following the word. The number indicates the accented mora; that is, the mora immediately following the accented mora that falls in pitch. In accentless words (about 80% of the Japanese lexicon), represented by "0", the first mora must be lowered by the above rule, and the rest remain high up to and including the following particle (such as the subject marker が /ga/). This is in contrast to words whose final mora is accented (尾高型 /odakagata/), such as /kagami/ (鏡 'mirror') (accent pattern "3"), in which the accent drops immediately after the word. That is, /katati ga/ (accentless) has a pitch pattern of LHH(H), whereas /kagami ga/ has a pattern of LHH(L).

Table 9: Accent Codes

Ortho-graphic	Kana	Phonetic	Accent	Pitch Pattern	Remarks
井川	イカワ	[ikawa]	1	HLL(L)	first mora accented
井田	イダ	[ida]	0	LH(H)	accentless
磯貝	イソガイ	[isoŋai]	2	LHLL(L)	second mora accented
鏡	カガミ	[kaŋami]	3	LHH(L)	last mora accented
形	カタチ	[katatɕi]	0	LHH(H)	accentless

5. JAPANESE PHONETIC DATABASE

The days of metallic, flat voices (as spoken by robots in science fiction movies) are over. Users are increasingly expecting natural speech from computers, not just properly pronounced, but also properly accented. This means that it is not only necessary to eliminate the phonemic ambiguities resulting from kana orthography, but also to pay close attention to the generation of allophonic variants and correct accent patterns.

Below is a brief description of the 130,000-entry Japanese Phonetic Database (JPD), developed in collaboration with The National Language Research Institute (国研) that provides, for the first time, IPA transcriptions as well as accent codes. Our research, including personal communications with several of Japan's leading speech technology experts, has confirmed that this is the first time such a database has been compiled [4].

The JPD was compiled to meet the needs of speech technology requirements, especially the generation of natural speech, by experienced editors trained in Japanese phonetics and phonology. The most important feature of this database is the hundreds of thousands of allophonic variants that include all the allophonic variation features described in this paper, such as vowel devoicing and pitch accent codes.

The principal methodology used to compile JPD is as follows:

1. In-depth analysis of Japanese allophonic variation and pitch accent patterns.
2. Development of a semi-automatic kana-to-IPA generator including major and minor allophones based on our comprehensive Japanese lexical database covering general vocabulary and proper nouns [5].
3. Compiling a database of ambiguous kana grapheme to phoneme mappings, such as 問う, in which the reading トウ represents [toɯ] and not the more common [toː].
4. Development of a semi-automatic accent code generation system.
5. Human proofreading of generated allophones and semiautomatic creation of accent codes.

The principal data fields in JPD are as follows:

1. Phonological/phonetic attributes, such as kana readings, disambiguated kana strings, IPA transcription and accent codes.
2. All major and many minor allophonic variants in IPA.

3. Grammatical information such as part-of-speech codes and conjugation patterns.
4. Semantic classification codes such as the type of proper noun (surname, male given name, etc.)

Table 10: JPD Sample Data

Headword	POS	TYPE	Accent	Reading	Phonetic	Remarks
鏡	NC	-	3	カガミ	kaŋami	voiced velar nasal
鏡	NC	-	3	カガミ	kagami	voiced velar stop
鏡	NC	-	3	カガミ	kaɣami	voiced velar fricative
自分	NC	-	0	ジブン	dʑibɯɴ	weakening of voiced alveolopalatal plosive (fricativized)
続く	V5	-	0	ツヅク	tsɯ̥zɯkɯ	devoicing of [tsɯ]
続く	V5	-	0	ツヅク	tsɯzɯkɯ	no devoiced vowel
恥	NC	-	2	ハジ	haʑi	voiced alveolopalatal fricative
恥	NC	-	2	ハジ	ha^{d}ʑi	voiced alveolopalatal affricate
本	NC	-	1	ホン	hoɴ	uvular nasalized consonant
本	NC	-	1	ホン	hõ	close front nasalized vowel
井上	NP	S	0	イノウエ	inoɯe	"ノウ" is pronounced to [oɯ], not long vowel
岸和田	NP	PS	0	キシワダ	ki̥ɕiwada	devoicing of [ki]
純子	NP	F	1	ジュンコ	dʑɯŋko	weakening of voiced alveolopalatal plosive （fricativized）
福生	NP	P	0	フッサ	ɸɯ̥s̚sa	first [s] is unreleased, devoicing of [ɸɯ]

6. FUTURE WORK

Speech technology systems, no matter how advanced or sophisticated, must have access to phonetic/phonological databases. To this end, our institute is engaged in research and development of CJK and Arabic phonetic databases that provide accurate transcriptions of how words are pronounced in actual speech. These databases are designed to advance Japanese speech technology, both TTS and ASR. Our future goals include compiling comprehensive IPA databases similar to JPD for Chinese, Korean and Arabic, as well as expanding JPD to cover more proper nouns.

7. REFERENCES

[1] Halpern, J. (2006) The Challenges of Intelligent Japanese Searching. Working paper (www.cjk.org/cjk/joa/joapaper.htm), The CJK Dictionary Institute, Saitama, Japan.

[2] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis”, Proc. of ICASSP, pp.1315-1318, 2000.

[3] Vance, T. J., *An Introduction to Japanese Phonology*, State University of New York Press, Albany, N.Y., 1987.

[4] Personal communication with Furui Laboratory (Tokyo Institute of Technology), the Advanced Telecommunications Research Institute International (Kyoto) and The National Language Research Institute (Tachikawa) (August/September 2008).

[5] Halpern, J., “The Role of Lexical Resources in CJK Natural Language Processing”, Proc. of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 64-71, COLING/ACL, Sydney, 2006.