Automatic Transformation of Environmental Sounds into Sound-Imitation Words Based on Japanese Syllable Structure

Kazushi Ishihara †, *Yasushi Tsubota* ‡, *Hiroshi G. Okuno* †

† Graduate School of Informatics, Kyoto University
‡ Academic Center for Computing and Media Studies, Kyoto University
{ishihara, tsubota, okuno}@kuis.kyoto-u.ac.jp

Abstract

Sound-imitation words, a sound-related subset of *onomatopoeia*, are important for computer-human interaction and automatic tagging of sound archives. The main problem of automatic recognition of sound-imitation word is that the literal representation of such words is dependent on listeners and influenced by a particular cultural history. Based on our preliminary experiments of such dependency and the sonority theory, we discovered that the process of transforming environmental sounds into syllable-structure expressions is mostly *listener-independent* while that of transforming syllable-structure expressions into sound-imitation words is mostly *listener-dependent* and influenced by culture. This paper focuses on the former lister-independent process and presents the three-stage architecture of automatic transformation of environmental sounds to sound-imitation words; segmenting sound signals to syllables, identifying syllable structure as mora, and recognizing mora as phonemes.

1. Introduction

The recent development of automatic speech recognition systems (ASR) has enhanced human-computer interaction and enabled speech input over normal or cellular phones. Current ASR's, however, fail in recognizing non-speech sounds, in particular environmental sounds such as animal voices, instrumental, natural, or machine sounds. Japanese speaking people often use *sound-imitation words*, a sound-related subset of onomatopoeia, *"giongo"*[1] in Japanese, to communicate such environmental sounds.

A sound-imitation word, or simply onomatopoeia in the rest of this paper, is a naming of a thing by a vocal imitation of the sound associated with it. For example, "w-a-N w-a-N" in Japanese and "bowwow" in English stand for the bark of a dog. However, "growl" is not a sound-imitation word, but a mimic word. Sound-imitation words are also a means of symbolic grounding as they transform sounds into a symbolic representation. In digital archives, sound-imitation words may be used for annotation such as in MPEG-7 for sound signals.

The literal representation of sound imitation-words is not unique but has a lot of variations that are depend on listeners. For example, "w-a-N w-a-N", "w-a-a-N w-a-a-N", "ky-a-N ky-a-N". This variation of literal representation is one of the main problems in automatic transformation of environmental sounds into sound-imitation words.

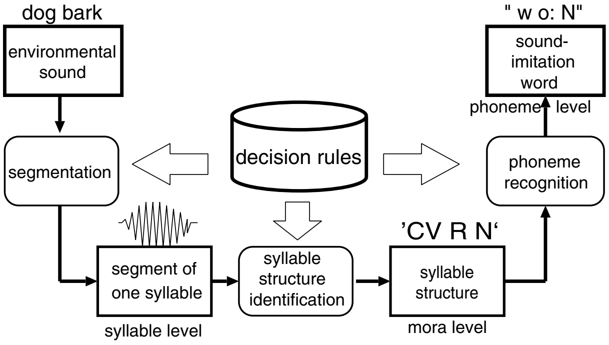

Figure 1: Sound-imitation word recognition processing

In this paper, we conjecture that such automatic transformation can be divided into listener-independent and -dependent processes. The process of transforming environmental sounds into syllable-structure expressions is mostly *listener-independent*, while that of transforming syllable-structure expressions into sound-imitation words is mostly *listener-dependent* and influenced by culture. This conjecture is verified by our preliminary experiments of such dependency and the sonority theory.

Section 2 describes the details of the preliminary experiments, Section 3 and 4 present the details of the listener-independent process that transforms environmental sounds into sound-imitation words by means of segmentation of sound signals to syllables, identifying syllable structure as mora, and recognizing mora as phonemes. These sections are devoted to dissolve ambiguities of allignment. Section 5 concludes the paper.

2. Listener-Dependency of Sound-Imitation Words

2.1. Three-stage method based on listener-independency

As mentioned above, the critical issue in the proposed transformation is how to resolve ambiguities in the literal representation. The conjecture of resolving this listener-independency problem is that some levels of representation in automatic sound-imitation word transformation should be listener-independent and some should be listener-dependent.

We assume that syllable- and mora-level expressions are listener-independent and phoneme-level expressions are listener-dependent. Based on this, we developed an approach to transform waveforms into sound-imitation words in three stages: (1) divide waveforms into one-syllable segments (syl-

This research is partly supported by the research grant from The Telecommunications Advancement Foundation.

[1] An action-related subset of onomatopoeia is called "gitaigo" or mimic words.

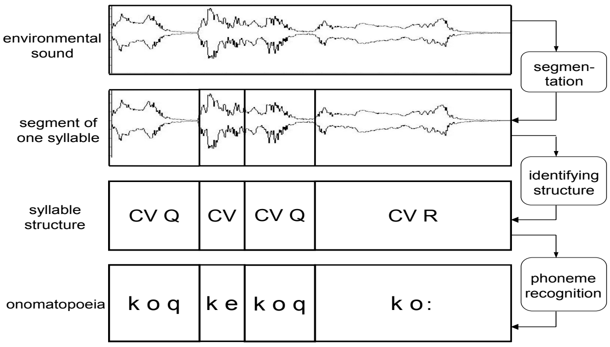

Figure 2: Recognition processing for a rooster's cry

Table 1: Japanese phonemes

vowels	a, e, i, o, u
consonants	b, by, ch, d, dy, f, g, gy, h, hy, j, k, ky, m, my n, ny, p, py, r, ry, s, sh, t, ts, w, y, z

lable level), (2) identify the syllable-structure of each segment (mora level), and (3) transform each segment from a syllable-structure to a sound-imitation word by using phoneme recognition (phoneme level) (Figure 1). A mora is a unit of spoken language that is longer than a phoneme and shorter than a syllable and usually consists of a consonant and a vowel in Japanese. The number of morae in a word is proportional to the sound length. Mora recognition is very important for Japanese language as it is mora-based, consisting of *kana*, a set of mora characters. The following explains each stage in the transformation.

Expressions at the syllable level are waveforms divided into one-syllable segments (Figure 2). Listener-independency at syllable level means that the number of syllables in representations by listeners agrees with and that the alignments in representations agree with. Listener-independency is confirmed by the sonority theory as described in Section 3.1. At the syllable level, environmental sounds are represented by the following regular expression: (syllable)+.

Expressions at the mora level are called syllable structures, meaning the structure of one syllable, and are sequences consisting of four kinds of Japanese mora symbols. These symbols are classified into normal and special morae. The following is a classification of Japanese mora symbols (Table 1).

- Normal morae (CV: a *kana* character except for N)
  - vowel
  - consonant + vowel
- Special morae
  - R: second mora of long vowel
  - Q: moraic silence when emphatic
  - N: moraic nasal

CV is one vowel and one consonant preceding the vowel. The consonant can be omitted. R, Q, and N are considered special mora symbols in studies of Japanese language and phonetics. N is a special consonant, because it is the only consonant that need not precede a vowel. Q is a silence followed by a glottal stop. It is counted as one mora, and it is the same length as CV based mora. At the mora level, a long vowel is two morae and R is the second mora in this case. A diphthong also consists of two morae but the difference between a diphthong and two vowels is ambiguous in the Japanese language, and thus we do not use a symbol to represent diphthongs. Table 2 shows the structures of one-syllable sound-imitation words in Japanese [1]. For Japanese speaking people that sound-imitation words of one syllable can be classified as in this table and that the expressions are listener-independent at mora level is obvious. Some studies are based on this conjecture [2, 3], however no theory has yet confirmed this. To test this conjecture, we have conducted perceptual experiments as described in Section 2.2.

Table 2: Structure of one-syllable onomatopoeia in Japanese

mora structure	instances of onomatopoeia
CV	hu, pi, gyo
CV R	sa:, su:, gu:
CV Q	kiq, saq, piq
CV R Q	za:q, do:q, ge:q
CV N	kaN, gaN, piN
CV R N	ga:N, go:N, bi:N

To transform sounds into sound-imitation words, we also need to perform recognition at the phoneme level. Recognition at this level, however, becomes listener-dependent and cannot identify the phoneme uniformly. This problem is dealt with in Section 4.2.

2.2. Preliminary Experiments

2.2.1. Method & benchmarks

To confirm the conjecture of listener-independency, we conducted two perceptual experiments. In the first, 19 subjects listened to 22 environmental sound stimuli – for example, animal voices or impact and instrumental sounds [S1, S2, S3] – and transcribed them freely. In the second experiment, 11 subjects listened to 24 environmental sound stimuli – mainly impact and friction sounds [S3, S4, S5, S6]– and transcribed them by applying structure rules, which minimize minute and inessential ambiguities. The subjects did not know what sounds the stimuli were.

2.2.2. Results and Observations

In the first experiment, about 65% of transcriptions agreed at syllable level and the number of syllables is equal to that of peaks in power envelope. For the syllable structure, the ratio of agreement was around 45%. In the second experiment, about 77% agreed at the syllable level, 74% at the mora level, and 50% at the phoneme level. Based on these results, we confirmed that the conjecture of listener-independency is correct.

3. Dividing into One-Syllable Segments

3.1. Segmentation based on the sonority theory

In the first step, waveforms are divided into one-syllable segments based on the results of the preliminary experiment. We developed a segmentation method according to the listener-independency conjecture that the number of syllables is equal to that of peaks in a power envelope. The method involves calculating the ratio of local minima between two peaks in a power envelope to the less one of two peaks and segmenting at the index of the local minima if the ratio is less than a threshold.

The listener-independency at the syllable level is confirmed by the sonority theory that syllabicity peaks coincide with sonority ones [5]. Sonority of a sound is its loudness relative to

Table 3: Segmentation accuracy

	proposed method	Julian	HMM
recall ratio	83.7 %	100.0 %	89.1 %
precision ratio	99.1 %	26.2 %	38.9 %

that of other sounds with the same length, stress, and pitch. The sonority of vowels is greater than that of consonants or semi-vowels, and that of voiced consonants is greater than that of voiceless ones. This theory confirms the listener-independency at syllable level, where listeners agree on the number of syllables in the majority of words. Several exceptions to this theory are evitable in Japanese because all consonants except for N must precede a vowel.

3.2. Segmentation experiment

We evaluated the effectiveness of the proposed method for environmental sound recognition. The same benchmark is also applied to an ASR system (Julian [4]) and an HMM-based system to compare to the proposed method. HMMs are 16-mixture monophone models constructed with 3795 environmental sound stimuli [S7, S8] labeled by five people. The model features are 26-dimension MFCC. The segmentation accuracy was evaluated by 157 syllable-boundaries of the benchmark. The results show the effectiveness of the proposed method (Table 3).

4. Transformation into Sound-Imitation Words

Here a rule-based approach for transforming each segment into a one-syllable sound-imitation word is introduced. Statistical methods such as the HMM are unreliable because the transformation system cannot possibly learn all existing environmental sounds due to the great variety. In fact, environmental sound recognition using HMM-based techniques constructed by environmental sound stimuli or using automatic speech recognition system is too different from the recognition by Japanese speaking people. For example, most Japanese speaking people listen to the call of a cock as "ko q ke ko q ko:", but an automatic speech recognition system recognizes it as "u a ga u ra a: a a ga ri ji hu: a u u u u u"(Figure 3). To resolve this issue, the syllable structure should be determined by features, that do not depend on what kind the sound stimuli are.

4.1. Identifying Syllable Structure

4.1.1. Syllable Structure

In the second step, the syllable structure of each segment is identified. From Table 2, a syllable is defined as the following structure [1]: CV [R] [N | Q]. Characters between square brackets can be omitted. The proposed structure is the same as that determined by Japanese grammar rules. One syllable generally has one vowel including diphthongs. This is confirmed by the sonority theory [5]. Therefore, a one-syllable structure can be represented as C* V C*, where C is a consonant and V is a vowel. * means closure. According to Japanese grammar rules, all consonants except for N must precede a vowel (N* C V N*). Additionally, N must follow or precede a vowel ([C] V N). We represent this structure by using mora symbols. Because a long vowel is two morae, V should be represented as V' [R]. V' is a vowel of one mora and R is the second mora of a long vowel. We use CV because C is not a mora symbol (CV [R] [N]). Finally, we adopt a special mora symbol, Q, for Japanese. As mentioned above, Q is one mora in Japanese. It cannot precede or follow N, precede R, or be at the beginning of a syllable. Consequently, the structure of one syllable is represented as CV [R] [N |Q]), which is the same as the above structure. Nonstandardized structures may appear in comics, but they are not used in general representation.

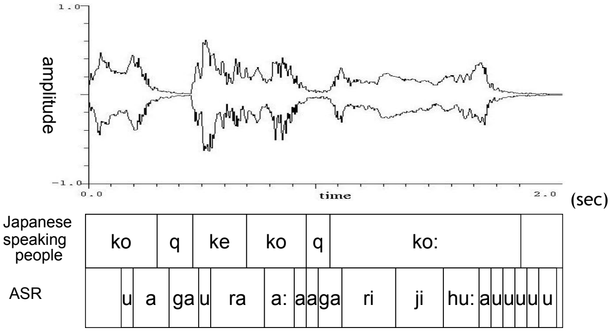

Figure 3: recognition result of rooster's crying

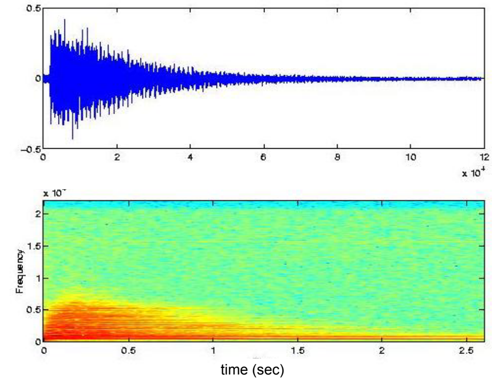

Figure 4: Environmental sound (N)

Determining R, Q, and N of a sound produces its syllable structure. Observation of waveforms and spectrograms of environmental sound stimuli used in preliminary experiments suggest that the length of the sound and the power decay rate determine the existence of R, Q and N (Figure 4). For example, in preliminary experiments, the ratio of samples containing R exceeded 80 % in a sample set where sample was longer than 400 ms, but it was 0 % in a sample set where the samples were shorter than 400 ms.

Studies of short-time sounds (called *tanpatsuon* in Japanese) have confirmed the conjecture. Short-time sounds are similar to segments to a point that a sound-imitation word of one syllable represents each sound. Tanaka [2] has claimed that a short-time sound with a long reverberation tends to be recognized as a sound-imitation word containing N. Hiyane [3] conducted cognitive experiments with short-time sounds generated by Gammatone. He claimed that reverberation up to 100 ms at 4 kHz was heard as 'chi', that between 100 ms and 200 ms as 'chi-n', and that above 300 ms as 'chi-i-n'. This claim agree with the results of our preliminary experiments.

4.1.2. Experiments to identify syllable structures

Identification is based on pre-computed decision trees that consider the length of the sound and the power decay rate of the

Table 4: Features to identify syllable structure

Range: from index of the maximum peak to index of MIN, which has the same power as 5 % of the power of the maximum peak
1. sound length
2. range length
3. average range gradients
4. maximum range gradients
5. maximum gradients of the lines drawn from MIN to each point in the range

Table 5: Accuracy of identifying syllable structure

	proposed	Julian	HMM
Q (recall ratio)	82.9 %	7.1 %	21.4 %
Q (precision ratio)	66.5 %	33.3 %	75.0 %
N (recall ratio)	22.7 %	9.1 %	9.1 %
N (precision ratio)	35.7 %	33.3 %	100.0 %
R (recall ratio)	84.6 %	40.0 %	66.7 %
R (precision ratio)	100.0 %	100.0 %	71.4 %

waveform (Table 4). We evaluated the effectiveness of the proposed method by testing the environmental sound recognition. The same benchmark is also applied to Julian and HMM. 25 samples are used to evaluate the accuracy of identifying the syllable structure. The results show the effectiveness without determining N (Table 5). The recognition rate of the proposed method was higher than that of Julian and the HMM-based system. The accuracy of determining N, however, was too low for the proposed method to be used as a recognition system. Japanese people determine N depending on the length of the sound and power decaying rate as well as on the frequency features and the type of sound. These issues need to be studied further.

4.2. Constructing phoneme-groups

Phoneme recognition is necessary to transform sounds from a syllable structure into sound-imitation words. However, phoneme-level recognition is listener-dependent. The conjecture to resolve this problem is that recognition at *phoneme-groups* level should be listener-independent. A phoneme-group is a set consisting of several phonemes that listeners in preliminary experiments tends to use often to represent the same sounds. An analysis of perceptual experiments has produced one phoneme classification (Table 6), which is used in constituting HMMs on 3.1 and 3.2. This phoneme classification is almost equal to a classification according to the manner of articulation. It means that phonemes articulated in the same manner tend are used to represent similar properties of environmental sounds. The fact is similar to the *sound-symbolism* theory, which is a study of the relationship between the sound of an utterance and its meaning [1]. The study claims that a phoneme itself suggests some meaning. For example, /m/ and /n/ (nasal consonant) tend to be used in soft expressions.

We plan to conduct perceptual experiments using this phoneme-group classification to develop more applicable classification. Additionally, we will design a method for phoneme-group recognition based on phoneme information such as pitch and sonority.

Table 6: Phoneme groups of Japanese consonants

1. m n my ny	(nasal)
2. j s sh z	(fricative)
3. f h hy	(fricative)
4. w y	(semi-vowel)
5. b by d dy g gy	(voiced plosive)
6. r ry	(liquid)
7. ch k ky p py t ts	(voiceless plosive etc.)

5. Conclusions

In this paper, we presented the three-stage architecture for automatic transformation of environmental sounds into sound-imitation words based on the conjecture of listener-independency at the syllable and mora levels. This conjecture is also verified by the experiment. We are implementing the three-stage architecture, of which performance will be reported as a separate paper. Another important future work is to recognize phoneme-groups by combining each segment with a one-syllable sound-imitation word (onomatopoeia). In addition, the model-driven automatic sound-imitation word should be investigated. Given the details of the sound source, such automatic transformation is adapted to the specific sound source. Ultimately, this automatic transformation system will be applied to enhance the sound-related human interaction system including ASR.

6. References

[1] Ikuhiro TAMORI, Lawrence Schourup: *Onomatopoeia* (in Japanese), Kuroshio Publisher, 1999.

[2] Kihachiro TANAKA: Study of Onomatopoeia Expressing Strange Sounds (Case of Impulse Sounds and Beat Sounds) (in Japanese), *Transactions of the Japan Society of Mechanical Engineers Series C*, Vol.61, No.592, 1995.

[3] K. Hiyane: Study of Spectrum Structure of Short-time Sounds and its Onomatopoeia Expression (in Japanese), *Technical Report of IEICE*, SP97-125, 1998.

[4] A. Lee, T. Kawahara, and K. Shikano: Julius – an open source real-time large vocabulary recognition engine. *Proc. EUROSPEECH*, 1691–1694, 2001.

[5] Peter Ladefoged: *A Course In Phonetics*, Harcourt Brace College Publishers, Captar10, 1993.

[S1] Tsuruhiko KABAYA and Michio MATSUDA, The Songs & Calls of 420 Birds in Japan (in Japanese), SHOGAKUKAN, 2001.

[S2] JUST BIRDS&ANIMALS, Sound Ideas, 2000.

[S3] *SHI-N KO-KA-O-N DA-I-ZE-N-SHU* (in Japanese), KING RECORD.

[S4] JURASSIC DINOSAURS, Sound Ideas, 2001.

[S5] IMPACT EFFECTS, Sound Ideas, 2000.

[S6] *KO-KA-O-N DA-I-ZE-N-SHU* (in Japanese), KING RECORD.

[S7] ANIMAL TRAX, Crypton Future Media, Inc.

[S8] WORLD WIDE OF ANIMALS, Crypton Future Media, Inc.