Consideration of Infants' Vocal Imitation Through Modeling Speech as Timbre-Based Melody

Nobuaki Minematsu[1] and Tazuko Nishimura[2]

[1] Graduate School of Engineering, The University of Tokyo,
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
mine@gavo.t.u-tokyo.ac.jp
[2] Graduate School of Medicine, The University of Tokyo,
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
nt-tazuko@ams.odn.ne.jp

Abstract. Infants acquire spoken language through hearing and imitating utterances mainly from their parents [1,2,3] but never imitate their parents' voices as they are. What in the voices do the infants imitate? Due to poor phonological awareness, it is difficult for them to decode an input utterance into a string of small linguistic units like phonemes [3,4,5,6], so it is also difficult for them to convert the individual units into sounds with their mouths. What then do infants acoustically imitate? Developmental psychology claims that they extract the holistic sound pattern of an input word, called *word Gestalt* [3,4,5], and reproduce it with their mouths. We address the question "What is the acoustic definition of word Gestalt?" [7] It has to be speaker-invariant because infants extract the same word Gestalt for a particular input word irrespective of the person speaking that word to them. Here, we aim to answer the above question by regarding speech as timbre-based melody that focuses on holistic and speaker-invariant contrastive features embedded in an utterance.

1 Introduction

Many speech sounds are produced as standing waves in a vocal tube, and their acoustic characteristics mainly depend on the shape of the tube. No two speakers have the same tube, and speech acoustics vary by speaker. Different shapes cause different resonances, which cause different timbre[1]. Similarly, different vowels are produced in a vocal tube by changing the tube's shape. Acoustically speaking, both differences between speakers and differences between vowels arise from the same cause. Speech features can also be changed by other factors such as features of a microphone, acoustics of a room, transmission characteristics of a line, auditory characteristics of a hearer, etc.

Despite the large acoustic variability, humans can accurately perceive speech. How is this done? Despite the progress of speech science, the contrast between the

[1] In musicology, timbre of a sound is defined as its spectral envelope pattern.

K. Satoh et al. (Eds.): JSAI 2007, LNAI 4914, pp. 26–39, 2008.

variability of speech acoustics and the invariance of speech perception remains an unsolved problem [8]. Speech engineering has attempted to solve this problem by collecting large numbers of samples corresponding to individual linguistic categories, e.g. phonemes, and modeling them statistically. IBM announced that they had collected samples from 350,000 speakers to build a speech recognizer [9]. However, no child needs this many samples to be able to understand speech. Perhaps the majority of the speech a child hears is from its mother and father. After it begins to talk, about a half of the speech a child hears is its own speech.

Developmental psychology explains that infants acquire spoken language by imitating the utterances of their parents. It is a tenet of anthropology that this behavior is found in no primates other than humans [2]. Further, we can say with certainty that infants never imitate the voices of their parents and that this is a clear difference from the vocal imitation of myna birds who imitate many sounds (cars, doors, animals, etc) as well as human voices. Hearing an adept myna bird say something, one can identify its owner [10]. However, the vocalizations of a child offer no clue as to the identity of its parents. What in the parents' voices do infants imitate? Due to poor phonological awareness, it is difficult for them to decode an input utterance into a string of phonemes [3,4,5,6], so it is also difficult to convert the individual phonemes into sounds. What then do infants imitate acoustically? Developmental psychology claims that they extract the holistic sound pattern of an input word, called *word Gestalt* [3,4,5], and reproduce it with their mouths. What then is the acoustic definition of word Gestalt? It must be speaker-invariant because infants can extract the same Gestalt irrespective of the person talking to them. To the best of our knowledge, no researcher has yet succeeded in defining it [7].

We recently formulated the above problem as a mathematical one and found a holistic and speaker-invariant representation of speech [11,12]. Here, an utterance is regarded as timbre-based melody. We did an experiment to test this representation. Acoustic models built with samples from only eight speakers based on the representation showed a slightly better recognition rate than Hidden Markov Models (HMMs) built with 4,130 speakers [13,14]. When we formulated the variability-invariance problem, we referred heavily to old and new findings in studies of linguistics, anthropology, neuroscience, psychology, language disorder, and musicology. Based on these findings and our results, we believe that infants extract the holistic and speaker-invariant contrastive features of speech.

2 Absolute Sense and Relative Sense of Sounds

2.1 Perception of Different Sounds as Identical

Figure 1 shows two utterances of /aiueo/, produced by two speakers. The one on the left was generated by a 200-cm-tall speaker and the one on the right was generated by an 80-cm-tall speaker. Although there is a large acoustic difference between the utterances, it can easily be perceived that they carry the same linguistic content, that is, /aiueo/. How do we perceive this equivalence in different stimuli? Do we perceive the equivalence after converting the two utterances into

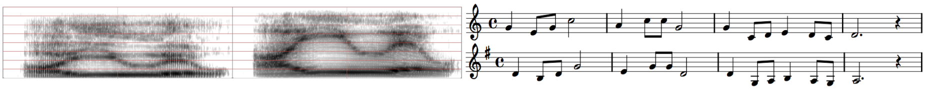

Fig. 1. Linguistic content of /aiueo/ uttered by two different speakers and the same piece of music played in two different keys

two phonemic sequences (sound symbol sequences) and finding the string-based equivalence between the two?

We think that the current speech recognition technology requires an answer of yes to the above question because the technology is based on a sound-to-symbol conversion technique that identifies separate sounds as single units among the linguistic sound categories, i.e. phonemes. However, this strategy dictates that acoustic models of the individual phonemes have to be made with samples from many speakers because a symbol corresponds to a variety of sounds.

Young children can also perceive the equivalence between the two utterances. Developmental psychology claims that infants do not have good phonemic awareness or a firm grasp of sound categories. This means that it is difficult for them to symbolize a separate sound and that invariance in perception is not due to string-based comparison. As explained in Section 1, infants first learn the holistic sound patterns in utterances and the individual phonemes later. This implies that invariance in perception must be based on comparison of holistic patterns. The question then is “What is the acoustic definition of the holistic and speaker-invariant pattern in speech?”

2.2 Relative Sense of Sounds in Music – Relative Pitch

Figure 1 also shows two pieces of the same melody performed in two different keys; C-major (top) and G-major (bottom). Although the acoustic substance of the two performances is very different, humans can easily perceive the equivalent musical content. When one asks a number of people to transcribe the two pieces as sequences of Do, Re, Mi, etc, three kinds of answers can be expected. Some will answer that the first one is So-Mi-So-Do La-Do-Do-So and the second one is Re-Ti-Re-So Mi-So-So-Re. These people are said to have absolute pitch (AP) and Do, Re, and Mi are pitch names for them, i.e. *fixed Do*. Others will claim that, for both pieces, they hear in their minds the same internal voices of So-Mi-So-Do La-Do-Do-So. They are said to have relative pitch (RP) and can verbalize a musical piece. For them, Do, Re, and Mi are syllable names, i.e. *movable Do*. The last group will not be able to transcribe the music, singing only “La-La-La-La La-La-La-La” for both. They also have RP but cannot verbalize a musical piece. They perceive the equivalence without sound-to-symbol conversion and only with a melody contour comparison. It should be noted that the RP people, the second and third groups, cannot identify a separate tone as one among the tonal categories.

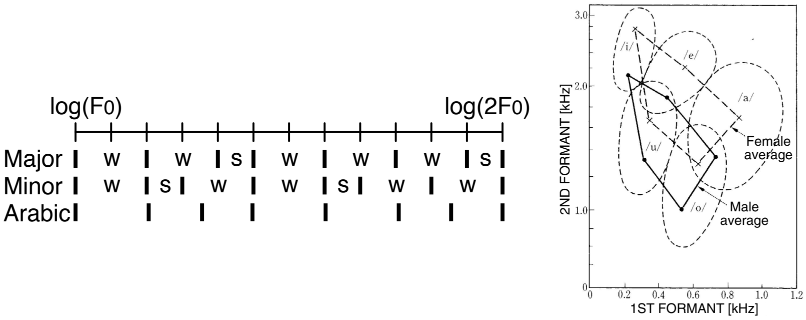

Fig. 2. Musical scales in octaves and a Japanese vowel chart using F_1 and F_2

AP people can memorize the absolute height of tones and use the heights to name musical tones. RP people capture the difference in the height between two tones (musical interval). If one explicitly defines the acoustic height of the Do sound, all the RP people, including the "La-La" people, can verbalize a given melody based on that definition. The difference between the second and third groups is that the former do not need a helper to define Do acoustically. How can they name the individual tones with no memory of the absolute height of tones? This ability is due to the tonal scale embedded in music, and, because this scale structure is key-invariant, the second group of people can easily identify the incoming tones independently of key.

Figure 2 shows three musical scales, all of which consist of octaves, eight tones in a group. The first two are well-known Western music scales, called major and minor. The third one is an Arabic music scale. For major and minor scales, an octave is divided into 12 semitone intervals, and eight tones are selected and arranged so that they have five whole tone intervals and two semitone ones. If C is used for tonic sound (the first sound) in a major scale, the key is called C-major and the tonal arrangement is invariant with key. The second group of people keeps the major and minor sound arrangements in memory and, based on these key-invariant arrangements, can identify the individual tones [15]. This is why they cannot symbolize a separate tone but can identify tones in a melody independently of key. Therefore, they find it difficult to transcribe a melody immediately after modulation in key. In contrast, people with AP can naturally transcribe the individual tones as pitch names even immediately after modulation. They sometimes do not notice the key change at all, but their identification is key-dependent, i.e., not robust at all.

RP people are capable of key-invariant and robust identification of tones because they dynamically capture the key-invariant sound arrangement [15]. In the following section, a similar mechanism is considered for speech perception.

2.3 Relative Sense of Sounds in Speech – Relative Timbre

A mother's voice is higher and a father's voice is lower because, in general, male vocal chords are heavier and longer [16]. A mother's voice is thinner and a father's

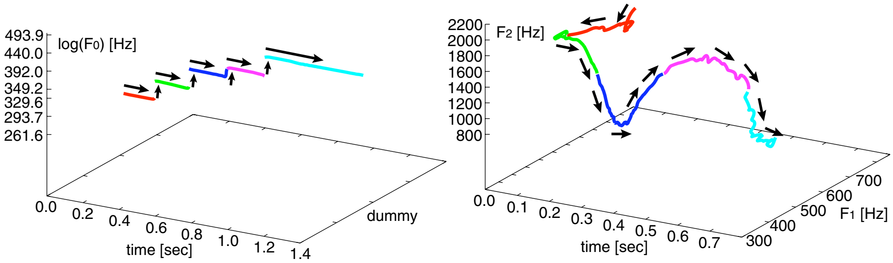

Fig. 3. Dynamic changes in pitch in CDEFG and changes in timbre in /aiueo/

voice is deeper because, in general, male vocal tracts are longer [16]. The former difference is one of pitch and the latter is one of timbre. The importance of RP is often discussed with regard to the former difference. People with strong AP tend to take a longer time to perceive the equivalence between a musical piece and a transposed version of it [17]. They often have to translate one symbol sequence consciously into another one to confirm the equivalence. Considering these facts, a similar mechanism, i.e. relative timbre, is hypothesized to explain why infants can perceive the equivalence between the two utterances in Figure 1 but cannot distinguish discrete symbols in the utterances.

As far as we know, however, all the discussions of sound identification in speech science and engineering have been based on absolute identification. How is it possible to discuss the relative identification of speech sounds based on invariant sound structure? Music consists of dynamic changes in pitch. Similarly, speech consists of dynamic changes in timbre. In Figure 3, a sound sequence of CDEFG played on a piano and a speech sound sequence of /aiueo/ are shown. Dynamic changes are visualized in a phase space. Pitch is a one-dimensional feature of F_0 and timbre is tentatively shown as a two-dimensional feature of F_1 and F_2. Cepstrum coefficients can also be used to expand the timbre space. Tonal transposition of a melody translates the dynamic changes in $F0$ but the shape of the dynamics is not altered. If the non-linguistic factors of speech such as speaker, microphone, etc, do not change the shape of the speech dynamics, the relative identification of speech sounds can be implemented on machines.

3 Robust and Structural Invariance Embedded in Speech

3.1 Mathematical Derivation of the Invariant Structure

As shown in the Japanese vowel chart in Figure 2, the male vowel structure can be translated into the female vowel structure. If the translation is accurate enough, the timbre dynamics can be easily formulated as invariant because differences in speakers do not change the sound arrangement and only transpose it multidimensionally. However, every speech engineer knows that this idea is too simple to be effectively applied to real speech data.

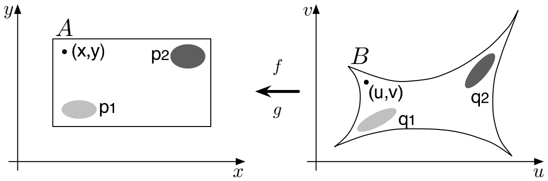

Fig. 4. Linear or non-linear mapping between two spaces

What kind of function can map the acoustic space of speaker A into that of speaker B: linear or non-linear? This question has been frequently raised in speaker adaptation research on speech recognition and speaker conversion research on speech synthesis. Figure 4 shows two acoustic spaces, one each for speakers A and B. Acoustic events p_1 and p_2 of A are transformed to q_1 and q_2 of B, respectively. If the two spaces have a one-to-one correspondence and point (x, y) in A is uniquely mapped to (u, v) in B and vice versa, transform-invariant features can be derived [18,13]. Every event is characterized as distribution:

$$1.0 = \oiint p_i(x,y)dxdy, \quad 1.0 = \oiint q_i(u,v)dudv. \tag{1}$$

We assume that $x{=}f(u, v)$ and $y{=}g(u, v)$, where f and g can be non-linear. Any integral operation in space A can be rewritten as its counterpart in B.

$$\iint \phi(x,y)dxdy = \iint \phi(f(u,v),g(u,v))|J(u,v)|dudv \tag{2}$$

$$= \iint \psi(u,v)dudv, \tag{3}$$

where $\psi(u, v) = \phi(f(u, v), g(u, v))|J(u, v)|$. $J(u, v)$ is Jacobian. Then, we get

$$q_i(u,v) = p_i(f(u,v),g(u,v))|J(u,v)|. \tag{4}$$

Physical properties of p_i are different from those of q_i. p_1 may represent /a/ of speaker A and q_1 may represent /a/ of B. We can show that the Bhattacharyya distance (BD) between two distributions is invariant with any kind of f or g.

$$BD(p_1,p_2) = -\log \oiint \sqrt{p_1(x,y)p_2(x,y)}dxdy \tag{5}$$

$$= -\log \oiint \sqrt{p_1(f(u,v),g(u,v))|J| \cdot p_2(f(u,v),g(u,v))|J|}dudv \tag{6}$$

$$= -\log \oiint \sqrt{q_1(u,v)q_2(u,v)}dudv \tag{7}$$

$$= BD(q_1,q_2) \tag{8}$$

The BD between two events in space A and that between their corresponding two events in space B cannot be changed. Substances can change easily, but their

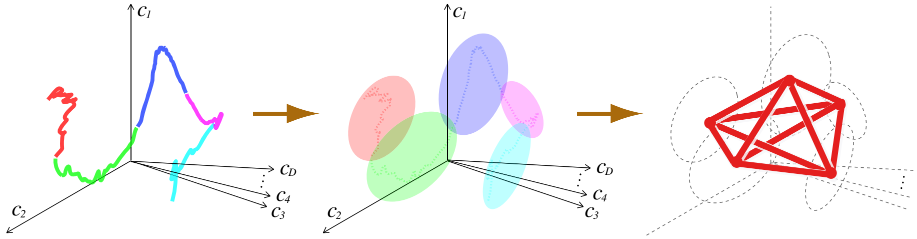

Fig. 5. Speaker-invariant structure in a cepstrum space without time axis

contrasts cannot be changed by any static transformation. The invariance is also satisfied with other distance measures such as the Kullback-Leibler distance. In this study, after some preliminary experiments, we adopted the BD.

The shape of a triangle is uniquely determined by fixing the lengths of all three sides. Similarly, the shape of an n point geometrical structure is uniquely determined if the lengths of all the ${}_{n}\mathrm{C}_{2}$ segments, including the diagonal ones, are given. In other words, if a distance matrix is given for n points, the matrix completely determines the shape of the n-point structure. As stated above, the BD is robustly transform-invariant. When n distributions are given, their BD-based distance matrix represents its robustly-invariant structure. An invariant structure can be extracted from an utterance. Figure 5 shows this procedure in a cepstrum space. After converting an utterance into a sequence of distributions, all the BD-based timbre contrasts between any two distributions are calculated.

3.2 Discussions of Structural Representation

Figure 6 shows Jakobson's geometrical system of French vowels [19]. He claimed that the same vowel system could be found irrespective of the speaker. It is well-known that Jakobson was inspired by the assertions of Saussure, the father of modern linguistics, who claimed that language is a system of conceptual and phonic differences and that the important thing in a word is not the sound alone but the phonic differences that make it possible to distinguish that word from all others [20]. The proposed invariant, holistic, and contrastive representation of an utterance can be regarded as a mathematical and physical interpretation of Saussure's claims and Jakobson's claims [11,12]. We discard sound substances and extract only phonic contrasts from an utterance because the former are very fragile and the latter are robustly invariant.

If Western music is played with the Arabic scale shown in Figure 2, it will take on a different color, i.e. Arabic accented Western music. This is also the case with speech. If the vowel arrangement of an utterance is changed, it will be a regionally accented pronunciation. Figure 6 also shows the vowel structures of two accented pronunciations of American English, plotted after vocal tract length normalization [21]. The vowel arrangement can change the color of pronunciation. There is good functional similarity between the sound structure in musical tones and that in vowel sounds. The difference may be just observations.

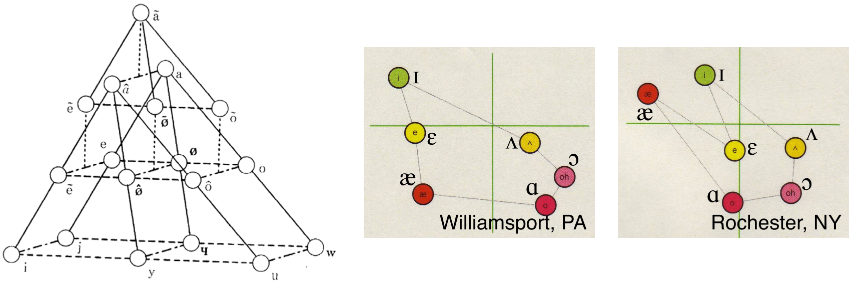

Fig. 6. Jakobson's geometrical structure of French vowels [19] and two accented pronunciations of American English [21]

Figure 3 shows a dynamic pattern or trajectory of pitch and of timbre. The pitch (melody) contour is often defined as a sequence of local pitch movements ($\Delta F0$), that is key-invariant. Similarly, the timbre contour can be defined as a sequence of local timbre movements (Δcepstrum), that is strongly speaker-dependent. It was mathematically and experimentally shown that vocal tract length differences change and rotate the direction of the timbre contour [22]. For example, with a frequency warping technique, a speech sample uttered by a male adult can be modified to sound like that of a boy. The warping operation shifts formant frequencies higher, that is, it makes them sound like the speaker is shorter. The direction of the Δcepstrum of a frame in the original speech and that of the corresponding frame in the modified speech was calculated. It was found that the timbre direction of a 170-cm-tall speaker and that of a 100-cm-tall speaker were approximately orthogonal. The directional difference became larger as the speaker's simulated height was increased or decreased. Further, the rotation of the timbre contour was not dependent on phonemes [22]. These results clearly indicate that the direction of local timbre dynamics is strongly dependent on the speaker. This is why, as shown in Figure 5, the directional components of the local dynamics are discarded and only the contrasts are extracted as scalar quantities. It should be noted that the proposed framework captures both local timbre contrasts and temporally distant contrasts.

4 Investigation Using Speech Recognition Research

4.1 Structural Acoustic Matching between Two Utterances

When two utterances are represented as different structures, how is the matching score between the two utterances to be calculated? As shown above, no transform can change the structure, which means that any transform can be interpreted as one of two geometrical operations, rotation or shift. As shown in Figure 7, the matching score for the two utterances is calculated as the minimum of the total distance between the corresponding two distributions (points in the figure) after shift and rotation. It was shown experimentally in [23] that minimum distance

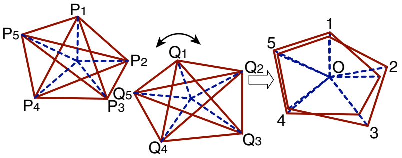

Fig. 7. Distance calculation

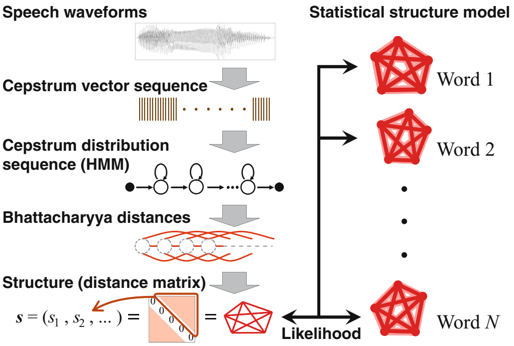

Fig. 8. Framework of structural recognition

D could be approximately calculated as a Euclidean distance between the two matrices, where the upper-triangle elements form the *structure vector*;

$$D(P,Q) = \sqrt{\frac{1}{n}\sum_{i<j}(P_{ij} - Q_{ij})^2}, \tag{9}$$

where P_{ij} is an (i,j) element of P and n is the number of distributions.

4.2 Verification of Structural Speech Recognition

To investigate the fundamental characteristics of the proposed framework, we examined automatic recognition of isolated words [13,14]. Here, a word was defined artificially as a connected vowel utterance. Since Japanese has the five vowels, /aiueo/, V_1-V_2-V_3-V_4-V_5 ($V_i{\neq}V_j$) utterances like /eoaui/ were used as words. The vocabulary size, i.e. perplexity, is ${}_5P_5$ (=120).

Eight male and eight female speakers recorded five utterances for each of the 120 words for a total of 9,600 utterances. Half of the samples from four males and four females were used for training and the others for testing. Since the proposed framework can eliminate differences between speakers well, only eight speakers were used for training. The recognition framework is shown in Figure 8. Vector sequences were converted into distribution sequences as MAP-based HMM training. Word templates were stored as statistical models averaged over structure vectors of each word's utterances.

As a comparison, an isolated word recognizer using tied-mixture triphone HMMs trained with 4,130 speakers [24] was built. Mel-frequencey cepstrum coefficients (MFCC) and its Δ were used with cepstral mean normalization (CMN). The word-based and vowel-based recognition rates are shown in Table 1. Although the proposed method completely discarded speech substances, it performed almost as well as the HMMs, which used both static and dynamic features. It should be noted that direct comparison is not fair because, for HMMs, the experiment was task-open, but, for the proposed framework, it was task-closed. In spite of this, we

Table 1. Recognition rates of the two methods

	HMM	Proposed
#speakers	4,130	8
word-based	97.4	98.3
vowel-based	98.8	99.3

can say that the proposed holistic and structural representation of speech identifies words well. Detailed descriptions of the recognition experiments are found in [13,14,25].

5 Investigation Using Speech Perception Research

5.1 Perception of Speaker-Variable and Size-Variable Speech

The RP people who can verbalize a given melody as a syllable name sequence have troubles transcribing it for some time immediately after the key has been modulated. We showed experimentally that this was also the case with speech [11]. Speaker-variable word utterances were generated with speech synthesis techniques and presented to human subjects. The speaker-variable utterances are those whose speaker information changes along the time axis. For example, the speaker is changed mora by mora or phoneme by phoneme. The presented stimuli were meaningless words. It was found that changing the speaker significantly degraded the transcription performance. Timbre changes due to speaker changes tended to be perceived as phoneme changes. However, the performance of a speech recognizer for the same stimuli was not degraded because the recognizer used speaker-independent HMMs trained with 4,130 speakers. A similar finding was obtained in another study [26] where size-variable speech samples were used.

In music, "La-La" people can enjoy music and can perceive the equivalence between a musical piece and a transposed version of it without symbolizing tones. We built a "La-La" machine, for which the two utterances in Figure 1 were completely identical but for which speech sound symbolization was impossible. We thought that this machine was a good simulator of infants' abilities and wondered whether this holistic speech processing could be found in adults.

5.2 Holistic Speech Processing Found in Adult Listeners

We found the answer in previous studies [27,28] done by another research group. Figure 2 shows a Japanese vowel chart. If vowel sounds of people the size of giants and those of people the size of fairies are obtained, they have to be plotted outside the ranges of existing people because formant frequencies depend on vocal tract length. Can subjects identify these vowel sounds in isolation? If they have difficulty with separate vowels, can they identify a continuous utterance more easily? The speaker-invariant holistic patterns are also embedded even in utterances of giants and fairies.

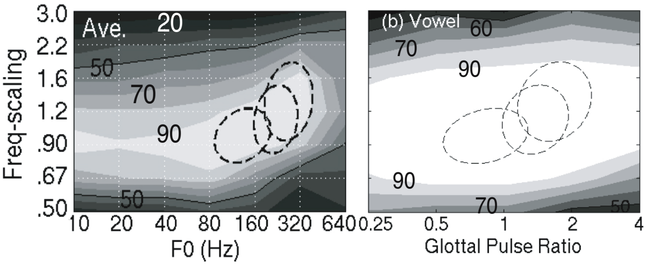

Fig. 9. Vowel identification rates without/with acoustic context [28]

Figure 9 (left) shows the identification rates of isolated Japanese vowels generated with various F_0s and body sizes. X and Y indicate F_0 and the ratio of frequency scaling in the spectrum, respectively. In the figure on the left, a vowel sample at (x, y) means that F_0 is x Hz and, roughly speaking, body height is $170/y$ cm ($5.6/y$ ft.). Three circles are the ranges of F_0 and body height for adult males, adult females, and children. Within these ranges, the identification rates are better than 90%. It is very difficult to absolutely identify separate vowels uttered by giants and fairies. For 65-cm-tall fairies (2.1 ft.) with an F_0 of 160 Hz, the performance is chance level (20%). Figure 9 (right) shows the identification rates of vowels in four-mora *unknown* words. Here, a vowel at (x, y) has an F_0 of $160x$ Hz. When giants and fairies say something in connected speech, subjects are reasonably able to identify the individual sounds even though the utterance is *meaningless*. For 65-cm-tall fairies, people can identify the sounds with about 60% accuracy. If known words are presented, performance will improve. With familiar words, it will improve drastically.

For “La-La” people, the request, “Remember the third tone in the next melody. Listen to another melody and raise your hand if you hear the same tone.” is very difficult to execute. Unless tones are symbolized, people will experience difficulty. Some people may have similar difficulty with the request, “Remember the third sound in the next utterance. Listen to another utterance and raise your hand if you hear the same sound.” Unless sounds are symbolized, people will have difficulty. To the best of our knowledge, two types of people have this kind of difficulty: young children and dyslexics. Their phonemic awareness is very weak [4]. Dyslexics are said to be good at seeing a whole tree but bad at seeing individual leaves [4,29]. However, both groups enjoy speech communication and can easily perceive the equivalence between the two utterances in Figure 1.

5.3 Non-robust Processing with Only Absolute Sense of Sounds

People with strong AP have some difficulty perceiving the equivalence between a musical piece and a transposed version of it [17]. Musicians with strong AP whose reference tone (A above middle C) is fixed at 440 Hz often have troubles performing music. An acoustic version of the reference tone depends on the orchestra playing it and it is sometimes 442 or 445 Hz. This small difference is difficult for AP musicians to accept. Absolute processing has to be non-robust.

If people have a strong absolute sense of speech sounds, they are likely to have difficulty perceiving the equivalence between the two utterances in Figure 1. Some autistics, who are considered to be good at seeing leaves but bad at seeing a whole tree [30,31], fall into this category. An autistic Japanese boy *wrote* that it was easy to understand his mother's speech but difficult to understand the speech of others [32]. However, it was also difficult for him to understand his mother's speech over the phone. He was able to write before he could speak, and spoken language was always difficult for him. Another autistic boy imitated voices as myna birds do, but spoken language was difficult also for him [34]. Autistics are much better at memorizing individual stimuli separately and absolutely as they are, but they are much worse at extracting holistic patterns hidden in the stimuli than non-autistic people [30,31]. It is explained in [30] that autistics lack the drive towards central coherence (Gestalt) and live in a fragmented world.

No child with normal hearing imitates voices, but myna birds and some autistics try to imitate voices as they are. Every speech synthesizer learns and imitates the voices of a single speaker. Every speech recognizer learns the voices of so many speakers by statistically modeling the acoustic features of the individual allophones separately and absolutely. However, the robustness in recognizing speech is far lower than that of humans. We cannot help considering the similarity between speech systems and autistics. In the 90's, AI researchers found that the robots sometimes behaved like autistics [33]. Both robots and autistics were bad at dealing with small environmental changes, known as the frame problem. AI researchers and therapists have recently been collaborating [33]. For them, making robots more suited to the real world and helping autistics become more accustomed to it are similar problems. Considering these facts, we think that speech engineers may have to face the same problem that AI researchers have.

6 Discussion and Conclusion

Anthropological studies showed that, basically speaking, no primates other than humans have relative pitch [35,36,37]. This is because relational processing requires a higher cognitive load than absolute processing. Therefore, other primates have difficulty perceiving the equivalence between the two musical pieces shown in Figure 1. Since pitch is one-dimensional and timbre is multi-dimensional, relative timbre processing should require an even higher load. What kind of behavior will be observed when the two utterances shown in Figure 1 are presented to chimpanzees? What if one of them is presented directly and the other is presented over the phone? If they cannot perceive the equivalence, human *spoken* language must also be very difficult. Researchers in the field of anthropology have made many attempts to teach human language to chimpanzees but most of them used visual tokens, not oral ones. Human vocal sounds failed to work as tokens even after being presented an enormous number of times [38]. However, young children can easily perceive the equivalence in different samples of speech.

We think that this invariant perception owes much to relative timbre, where an utterance is perceived as a timbre-based melody.

Finally, we want to carry out a thought experiment. Suppose that the parents of identical twins get divorced immediately after the twins are born and that one twin is taken in by the mother and the other is taken in by the father. What kind of pronunciation will the twins have acquired after ten years? The twins do not produce voices that sound, respectively, like the mother and like the father. However, there is an exceptional case in which the twins' pronunciations will be very different: when the parents are speakers of different regional accents. Timbre difference based on difference in speakers does not affect the pronunciation but that based on regional accents does. Why? The simplest explanation is that infants do not learn the sounds as they are but learn the sound system embedded in spoken language. The proposed representation extracts the invariant system embedded in an utterance. We believe that this is the answer to the question.

References

1. Kuhl, P.K., Meltzoff, A.N.: Infant vocalizations in response to speech: Vocal imitation and developmental change. J. Acoust. Soc. Am. 100(4), 2425–2438 (1996)
2. Gruhn, W.: The audio-vocal system in sound perception and learning of language and music. In: Proc. Int. Conf. on language and music as cognitive systems (2006)
3. Hayakawa, M.: Language acquisition and matherese. In: Language, Taishukan pub. vol. 35(9), pp. 62–67 (2006)
4. Shaywitz, S.E.: Overcoming dyslexia, Random House (2005)
5. Kato, M.: Phonological development and its disorders. J. Communication Disorders 20(2), 84–85 (2003)
6. Hara, K.: Phonological disorders and phonological awareness in children. J. Communication Disorders 20(2), 98–102 (2003)
7. Minematsu, N., Nishimura, T.: Universal and invariant representation of speech, CD-ROM of Int. Conf. Infant Study (2006), `http://www.gavo.t.u-tokyo.ac.jp/~mine/paper/PDF/2006/ICIS_t2006-6.pdf`
8. Johnson, K., Mullennix, J.W.: Talker variability in speech processing. Academic Press, London (1997)
9. `http://tepia.or.jp/archive/12th/pdf/viavoice.pdf`
10. Miyamoto, K.: Making voices and watching voices. Morikawa Pub. (1995)
11. Minematsu, N., et al.: Theorem of the invariant structure and its derivation of speech Gestalt. In: Proc. ISCA Int. Workshop on Speech Recognition and Intrinsic Variation, pp. 47–52 (2006)
12. Minematsu, N.: Are learners myna birds to the averaged distributions of native speaker? – a note of warning from a serious speech engineer –, CD-ROM of ISCA Int. Workshop on Speech and Language Technology in Education (2007)
13. Asakawa, S., Minematsu, N., Hirose, K.: Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics. In: Proc. InterSpeech, pp. 890–893 (2007)
14. Qiao, Y., Asakawa, S., Minematsu, N.: Random discriminant structure analysis for continous Japanese vowel recognition. In: Proc. Int. Workshop on Automatic Speech Recognition and Understanding, December 2007 (to appear)

15. Taniguchi, T.: Sounds become music in mind – Introduction to music psychology –. Kitaoji Pub. (2000)
16. Titze, I.R.: Principles of voice production. Prentice-Hall Inc., Englewood Cliffs (1994)
17. Miyazaki, K.: How well do we understand absolute pitch? J. Acoust. Soc. Jpn. 60(11), 682–688 (2004)
18. Minematsu, N., Asakawa, S., Hirose, K.: Linear and non-linear transformation invariant representation of information and its use for acoustic modeling of speech. In: Proc. Spring Meeting Acoust. Soc. Jpn., pp. 147–148 (2007)
19. Jakobson, R., Lotz, J.: Notes on the French phonemic pattern, Hunter (1949)
20. Saussure, F.: Cours de linguistique general. In: Publie par Charles Bally et Albert Schehaye avec la collaboration de Albert Riedlinge, Lausanne et Paris, Payot (1916)
21. Labov, W., Ash, W., Boberg, C.: Atlas of North American English. Walter de Gruyter, Berlin (2001)
22. Saito, D., et al.: Derectional dependency of cepstrum on vocal tract length. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing (2008, submitted)
23. Minematsu, N.: Yet another acoustic representation of speech. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp. 585–588 (2004)
24. Kawahara, T., et al.: Recent progress of open-source LVCSR engine Julius and Japanese model repository. In: Proc. Int. Conf. Spoken Language Processing, pp. 3069–3072 (2004)
25. Asakawa, S., Minematsu, N., Hirose, K.: Multi-stream parameterization for structural speech recognition. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing (2008, submitted)
26. Takeshima, C., Tsuzaki, M., Irino, T.: Identification of size-modulated vowel sequences and temporal characteristics of the size extraction process, IEIEC Technical Report, SP2006-29, 13-17 (2006)
27. Smith, D.R., et al.: The processing and perception of size information in speech sounds. J. Acoust. Soc. Am. 171(1), 305–318 (2005)
28. Hayashi, Y., et al.: Comparison of perceptual characteristics of scaled vowels and words. In: Proc. Spring Meeting Acoust. Soc. Jpn., pp. 473–474 (2007)
29. Davis, R.D., Braun, E.M.: The gift of dyslexia, Perigee Trade (1997)
30. Frith, U.: Autism: Explaining the enigma. Blackwell Pub., Malden (1992)
31. Happe, F.: Autism: An introduction of psychological theory. UCL Press (1994)
32. Higashida, N., Higashida, M.: Messages to all my colleagues living on the planet. Escor Pub. (2005)
33. Nade, J.: The developing child with autism: evidences, speculations and vexed questions. In: Tutorial Session of IEEE Int. Conf. Development and Learning (2005)
34. Asami, T.: A book on my son, Hiroshi, Nakagawa Pub., vol. 5 (2006)
35. Trehub, S.E.: The developmental origins of musicality. Nature neurosciences 6, 669–673 (2003)
36. Hauser, M.D., McDermott, J.: The evolution of the music faculty: A comparative perspective. Nature neurosciences 6, 663–668 (2003)
37. Levitin, D.J., Rogers, S.E.: Absolute pitch: perception, coding, and controversies. Trends in Cognitive Sciences 9(1), 26–33 (2005)
38. Kojima, S.: A search for the origins of human speech: Auditory and vocal functions of the chimpanzee. Trans Pacific Press (2003)