A Consideration of ASR Based on Animal Evolution and Human Development — What Should A of ASR Stand for? —

Nobuaki Minematsu

Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan

mine@gavo.t.u-tokyo.ac.jp

Abstract

Speech features are inevitably changed by static biases of extra-linguisitc factors, such as age, gender, microphone, etc. In the conventional ASR, these changes were often handled by building speaker-/environment-independent acoustic models trained with thousands of utterances produced in different conditions. Here, absolute properties of speech (speech substances) such as spectral envelopes were extracted and modeled statistically. Recently, contrast-based speech modeling has been proposed [1], where only speaker-invariant speech contrasts or dynamics (relative properties) are extracted and modeled. In this paper, firstly, these two models (substance or contrast) are compared through a discussion on how animals had acquired the ability of robust processing of stimuli and how animals still differ from humans. After experimental results of our proposed framework of ASR are shown, it is also discussed that the strategy of speech processing based on the conventional model is similar to that of severely damaged autistics. They have a good and exact memory of stimuli but are weak in handling changes of stimuli.

Index Terms: extra-linguistic features, speech contrasts, invariance, speech structures, automatic speech recognition, autistics

1. Nature of perceptual constancy

All the living systems receive stimuli from the external environment and generate some responses to it. Through this reception and generation loop, interaction emerges, where the same stimulus often changes in its shape and form. For example, a visual image is modified in its shape by viewpoint changes but our perception is constant and invariant. As for color, a flower in broad daylight and the same one at sunset give us different color patterns but we perceive the equivalence between them. Humming by a male and that of the same melody by a female often differ in fundamental frequency but we easily perceive the equivalence. This is the case with speech. Male voices are deeper in timbre than female ones but the invariant perception is easy between a father's "hello!" and a mother's. Although the above stimuli are given to receivers using different media, all the changes are caused commonly by inevitable static biases.

It seems that researchers of psychology found that a similar mechanism is working to cancel the static biases and realize the invariant perception [2, 3, 4]. Figure 1 shows the look of the same Rubik's cube seen through differently colored glasses. Although the corresponding tiles of the two cubes have different colors absolutely, we name them using the same labels. On the other hand, although we see four *blue* tiles on the top of the left cube and seven *yellow* tiles on the right, when their surrounding tiles are hidden, we see that they have the same color (See Figure 2). Absolutely different colors are perceived as identical and absolutely identical colors are perceived as different.

Similar phenomena are easily found in tone perception. Figure 3 shows two sequences of musical notes. The upper corresponds to humming by a female and the other to that of the same melody

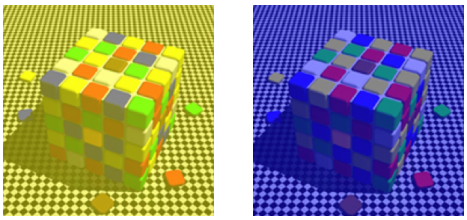

Figure 1: The same Rubik's cube seen with two colored glasses

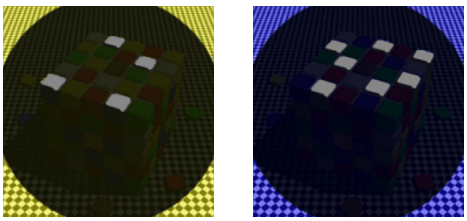

Figure 2: Perception of colors without context

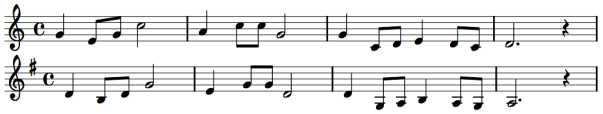

Figure 3: A musical melody and its transposed version

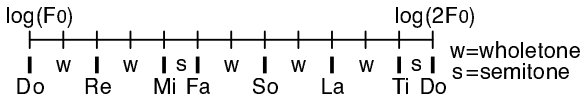

Figure 4: Tonal arrangement (scale) of the major key

by a male. If hearers have relative pitch and can transcribe these melodies, they convert the two melodies into the same sequence of syllable names (So Mi So Do La Do Do So). The first tone of the upper melody and that of the lower are different absolutely but they name these tones using the same label. The first tone of the upper and the fourth of the lower are identical absolutely but they claim that the two tones are different. Similar to the color perception, absolutely different tones are perceived as identical and absolutely identical tones are perceived as different.

Researchers found that the invariant perception, colors and tones, occurs mainly based on contrast-based information processing [2, 3, 4]. In other words, our invariant perception of colors and tones is guaranteed by the invariant relation of the focused stimulus to its surrounding stimuli. For individuals with relative pitch, a single tone is difficult to name but tones in a melody are easy to transcribe. If two tones in a melody of the major key, which can be temporally distant, are three wholetones apart in pitch, they must be Fa and Ti according to the tonal arrangement (scale) of the major key (See Figure 4). This arrangement is invariant with key and, using this arrangement as constraint, the key-invariant tone identification can occur.

As was found in ecology, the invariant color perception occurs even to butterflies and bees [5]. This color perception is extremely old evolutionarily. On the other hand, researchers of

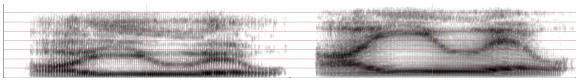

Figure 5: /aiueo/s produced by a tall speaker and a short one

anthropology found that the invariant tone perception is difficult even for monkeys [6]. What they claim is not that monkeys cannot transcribe a melody but that monkeys cannot perceive the equivalence between a melody and its transposed version [6]. The relative pitch perception is very new evolutionarily.

2. Human development of spoken language

How can infants acquire the ability of robust speech processing? Recently, in the field of AI, there is a research trend to focus on infants' acquisition and development of cognitive abilities [7, 8]. One obvious fact is that a major part of the utterances an infant hears are from its parents. After it begins to talk, about a half of the utterances it hears are its own speech. It can be said that the utterances an individual hears are strongly speaker-biased unless he/she has speaking disabilities. The variability problem of speech acoustics should be solved not by collecting samples if one wants to realize a human-like speech processor.

Infants acquire spoken language through imitating their parents' utterances actively, called vocal imitation. But they don't impersonate their parents. A question is raised; what acoustic aspect of the voices do infants imitate? One may claim that infants decompose an utterance into a phoneme sequence and each phoneme is realized acoustically by their mouths. But researchers of infant studies deny this claim because infants don't have good phonemic awareness [9, 10]. Then, what is imitated?

An answer from infant studies is the holistic sound pattern embedded in an utterance [9, 10], called otherwise as word Gestalt [11] and related spectral patterns [12]. The holistic pattern has to be speaker-invariant because, whoever speaks a specific word to an infant, its responses of imitation are similar acoustically. Another question is raised; what is the physical definition of the speaker-invariant holistic speech pattern?

The vocal imitation is rare in animals [13] and non-human primates scarcely imitate others' utterances [14]. This performance can be found in only a few species of animals, i.e. birds, whales, and dolphins. But there is a critical difference between humans and animals. Animals' imitation is basically the imitation of sounds like impersonation [13]. Take myna birds for example. They imitate the sounds of cars, dogs as well as human voices. Hearing a very good myna bird say something, one can guess its human owner [15] but cannot guess the parents of an infant by hearing its voices. Considering that the same pitch contours (intonation patterns) of different keys (genders or speakers) are equivalent for humans but different for monkeys and that the same linguistic contents acoustically generated by different speakers are equivalent for humans but probably different for animals [16], the ability of extracting an invariant pattern from a variable sound stream might be unique to humans.

3. Natural solution of speaker variability

As for speech, changes in vocal tract shape and length cause timbre changes. Basically speaking, dynamic and morphological changes of the vocal tract generate different phonemes acoustically. Static differences of the vocal tract length among speakers generate speaker variability acoustically. Figure 5 shows the same linguistic content generated by a tall speaker and a short one. What is the speaker-invariant holistic speech pattern?

Speaker difference is often modeled mathematically as space mapping in studies of voice conversion. This means that if we can find some transform-invariant features, they can be used as speaker-invariant features. Recently, some proposals have been done [17, 18, 19] but speaker variability was always modeled simply as $\hat{f}=\alpha f$ (f=frequency, α=constant). Many studies of speaker conversion adopted more sophisticated transforms, indicating that $\hat{f}=\alpha f$ cannot characterize speaker variability well enough. Further, it should be noted that all of these proposals tried to find invariant features in individual speech sounds, not in a holistic pattern only composed of speech contrasts.

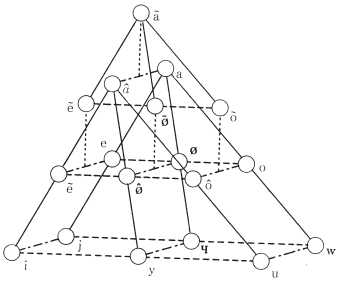

Figure 6: The invariant system of French vowels

As shown in [5], the perceptual constancy of colors is found in butterflies. But no researcher claims that a butterfly has statistical models of individual colors which are acquired by looking at the colors through thousands of differently colored glasses. Further, naming individual colors (elements) is not needed to perceive the equivalence between a flower in broad daylight and the same one at sunset. In contrast in ASR, acoustic and statistical modeling of individual phonemes (elements) using thousands of speakers (differently shaped tubes) is the most popular approach. From an ecological and evolutionary viewpoint, this strategy is remarkably weird and the invariant speech perception should be implemented on machines based on processing holistic patterns composed of invariant contrasts or relations.

A similar claim can be found in classical linguistics [20]. Jakobson proposed a theory of acoustic and relational invariance, called distinctive feature theory. He repeatedly emphasizes the importance of relational and systemic invariance among speech sounds by referring to phrases of other scholars such as Klein (topologist), Baudouin, and Sapir (linguists). Figure 6 shows his invariant system of French vowels and semi-vowels. Figure 4 is the key-invariant tonal arrangement in melody and Figure 6 is the speaker-invariant timbre arrangement in vowel sounds. Considering that pitch is one-dimensional but timbre is multi-dimensional, what has to be implemented on machines is a mechanism of relative timbre perception, where invariant and multi-dimensional timbre contrasts are used to determine the value of individual sounds. In a classical study of phonetics, the importance of relational invariance was verified in word identification tests [21]. It is interesting that Lagefoged discussed a good similarity between perception of vowels and that of colors.

Recently, a big European project has started to develop computational mechanisms needed to create artificial agents that can acquire verbal communication skills [7]. In this project, memory-prediction theory is adopted as its core framework. This theory was proposed by Hawkins, who made a strong claim as neuroscientist in his book [22], "I believe a similar abstraction of form is occurring throughout the cortex. Memories are stored in a form that captures the essence of relationships, not the details of the moment. The cortex takes the detailed, highly specific input and converts it to an invariant form. Memory storage and recall occur at the level of invariant forms."

4. Mathematical solution of the variability

4.1. Mathematically guaranteed topological invariance

Are there any invariant and contrastive features (measures) with respect to any linear or non-linear invertible transforms? In [23], we proved that f-divergence between two distributions is invariant with any kind of invertible and differentiable transforms (sufficiency). We also proved that any completely invariant measure with respect to two distributions has to be written in the form of f-divergence (necessity), which is formulated as

$$f_{div}(p_1, p_2) = \oint p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x}. \quad (1)$$

Figure 7 shows two spaces (shapes) which are deformed into each other through an invertible and differentiable transform. An event is described not as point but as distribution. Two events of p_1 and p_2 in A are transformed into P_1 and P_2 in B. The invariance of f-divegence is always satisfied [23].

$$f_{div}(p_1, p_2) \equiv f_{div}(P_1, P_2) \quad (2)$$

In a series of our previous studies [1, 23, 25, 26], we have been using Bhattacharyya distance (BD) as one of the f-divergence measures. Figure 8 shows a procedure of representing an input utterance only by BD. The utterance in a feature space is a sequence of feature vectors and it is converted into a sequence of distributions through automatic segmentation. Here, any speech event is modeled as a distribution. Then, the BDs are calculated from any pair of distributions to form a BD-based invariant distance matrix. As a distance matrix can fix a unique geometrical shape, we call the matrix as speech structure. Individual speech sounds can change but their entire system cannot change at all.

4.2. Some experimental results of isolated word recognition

Figure 9 shows the basic framework of isolated word recognition based on speech structures. To convert an utterance into a distribution sequence, the MAP(Maximum A Posteriori)-based HMM training is adopted. Then, the BD between any pair of the distributions is obtained. After calculating the structure, a structure vector is formed by using all the elements in the upper triangle. This vector is a holistic and speaker-invariant representation of a word utterance. The right-hand side of the figure shows an inventory of word-based statistical structure models for the entire vocabulary. The candidate word showing the maximum likelihood score is a result of recognition.

The speech structure is invariant with any kind of invertible transforms. This indicates that two different words can be evaluated as the same. To solve this problem, we introduced good constraints called Multiple Stream Structuralization (MSS) [25] so that we could obtain the invariance only with speaker variability. Due to the limit of space, MSS is not explained in details in this paper but interested readers should refer to [25, 26].

In [25, 26], structure-based isolated word recognition was compared to substance-based word recognition. The former used the proposed structure (contrast) models and the latter used the conventional word HMMs trained with spectrum-based (substance-based) features. Two word sets were used. In a set, a word was artificially composed of five vowels such as /eauoi/ and /uoaie/. As Japanese has only five vowels, PP=120. The other set was a Japanese phoneme-balanced word set and PP=220 [27]. To investigate the robustness with respect to mismatch between training and testing conditions, frequency warping was applied to testing samples to simulate speech samples generated by very tall and very short speakers. Table 1 summarizes the results.

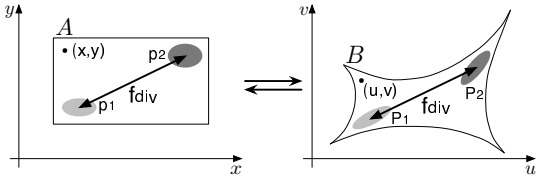

Figure 7: Topological deformation of manifolds (shapes)

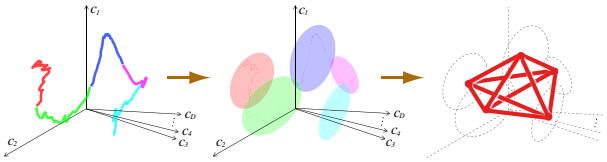

Figure 8: An utterance structure composed only of f-divergence

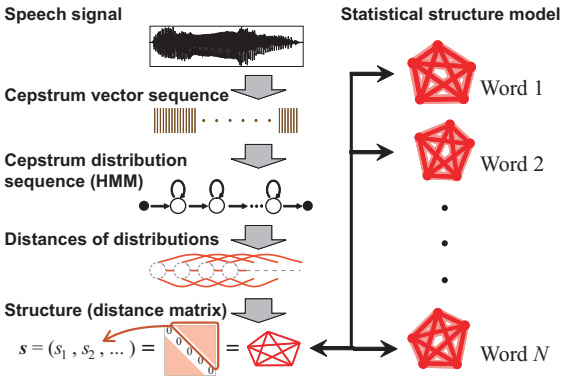

Figure 9: Framework of structure-based word recognition

Table 1: Comparison between HMMs and structures

(a) Results for five-vowel words

α	-0.40	-0.35	-0.30	-0.25	-0.20	-0.15	-0.10	-0.05	0.00
HMMs	0.92	0.94	1.75	6.83	21.8	40.5	60.2	80.0	83.9
matched	58.9	62.1	64.3	68.5	74.3	78.3	81.5	83.5	83.9
Structures	53.6	61.9	68.3	74.3	80.1	84.0	86.9	88.8	89.1

α	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
HMMs	78.2	63.1	44.5	24.8	8.85	1.88	1.00	0.67
matched	84.7	85.8	86.3	86.3	86.3	86.4	87.2	86.6
Structures	89.5	89.8	90.5	90.6	90.9	91.0	91.2	91.3

(b) Results for phoneme-balanced words

α	-0.40	-0.35	-0.30	-0.25	-0.20	-0.15	-0.10	-0.05	0.00
HMMs	5.33	11.2	21.5	37.4	57.6	74.1	87.6	95.8	98.3
matched	94.9	96.4	96.6	97.4	97.8	98.0	97.9	98.3	98.3
Structures	46.7	55.3	63.1	69.9	77.4	83.2	88.0	91.6	92.6

α	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
HMMs	97.5	92.3	81.2	64.6	45.6	27.2	14.0	7.65
matched	98.4	98.5	98.4	98.5	98.5	98.3	98.4	98.6
Structures	92.1	90.6	86.3	81.0	74.0	66.9	58.0	49.3

α is a warping parameter and varied from -0.4 to 0.4 at 17 steps. α=-0.4/+0.4 means doubling/halving the vocal tract length. Both HMMs and structures used no adaptation technique. The number of distributions per word is 25 for the vowel words and 30 for the balanced words. Matched shows the results of using 17 sets of matched conditioned HMMs. In the vowel word set, a single set of structures show almost the same or higher performance compared to the 17 matched HMM sets. In the phoneme balanced set, however, the performance of the structures is lower than that of HMMs at α=0.0 although the robustness of the structures is shown at $|\alpha|$>0.15. This is considered to be because unvoiced consonant sounds are less speaker-dependent and absolute features are needed for these sounds. Currently, we're integrating both the models for compensation [24]. Detailed description of the experiments is in [25, 26].

5. Discussions and conclusions

In this paper, the theoretical and underling reasons why we proposed a new representation of speech [1] are described in details. In a word, the conventional strategy of speech modeling is extremely unnatural considering animal evolution and human development of spoken language. To conclude this paper, a radical discussion is done on what A of ASR should stand for.

We can find individuals with a certain cognitive disorder, among whom, the following behaviors are observed. Only by looking at Figure 1, they can find that the four *blue* tiles on the left and the seven *yellow* tiles on the right have the same color [28]. A melody and its transposed version are just different sequences of tones [29]. Utterances of their own mothers are easy to transcribe but those of others are difficult [30]. But utterances of the mothers turn difficult on a telephone [30]. Their vocal imitation is basically impersonation like myna birds [31]. They are autistics and usually have an extremely good memory for the detailed, highly specific aspects of stimuli [16]. Perception of autistics is so different from that of normally developed individuals that not a few autistics describe themselves as aliens born on this planet [16, 30, 32] to explain how their perception is different. It is well-known that spoken language is very difficult for severely damaged autistics to use. Printed language, not spoken language, often becomes the first language.

Autistic professor of animal science, Temple Grandin, explains that the strategy of information processing of autistics is similar to that of animals [16]. It is local, concrete, and specific. For her, the strategy of normally developed individuals is holistic, abstract, and general. In a medical study [33], monkeys were used as models of severely damaged autistics.

As discussed in Section 2, although normal infants initially capture the detailed and concrete aspects of stimuli [20], they soon ignore some aspects and, for example, become able to find an invariant sound pattern in acoustically different but linguistically identical utterances. Considering the findings of evolutionary anthropology [6, 14], it can be reasonably hypothesized that only humans have a good abstraction ability to cancel static biases of pitch and timbre from auditory stimuli. It should be reminded that this abstraction should be based on holistic pattern processing if we consider how animals had acquired the ability of robust processing of visual stimuli. As claimed by researchers of infant studies, young children communicate orally to others with reduced phonemic awareness [9]. Similar performances are found in phonological dyslexics [10]. Both have difficulty in manipulating phonemes in utterances and then, in manipulating written language. But they have no trouble in oral communication. The structure-based speech recognizer cannot handle phonemes but their oral performance is good and robust.

For several decades, speech engineering has proposed methods of acoustically detailed modeling of utterances for speech synthesis and has refined methods of acoustic modeling of the individual phonemes (elements) for speech recognition. Considering human development of spoken language, I have to claim again that this strategy is extremely weird if speech engineers try to build human-like speech processors. If the goal of speech engineering is just providing text-to-speech and speech-to-text media converters for users, however, the internal mechanism of the converters does not have to fit to the human mechanism.

What should A of ASR stand for, then? A definitely clear and critical difference between normally developed individuals and the current speech recognizers forces me to claim that A of the current ASR has to stand for, not Automatic, but Autistic, Animal, or Alien. Where is the goal of speech engineering?

6. References

[1] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," *Proc. ICASSP*, 889–892, 2005.

[2] R. B. Lotto *et al.*, "An empirical explanation of color contrast," *Proc. the National Academy of Science USA*, 97, 12834–12839, 2000

[3] R. B. Lotto *et al.*, "The effects of color on brightness," *Nature neuroscience*, 2, 11, 1010–1014, 1999

[4] T. Taniguchi, *Sounds become music in mind – introduction to music psychology –,* Kitaoji Pub., 2000

[5] A. D. Briscoe *et al.*, "The evolution of color vision in insects," *Annual review of entomology*, 46, 471–510, 2001

[6] M. D. Hauser *et al.*, "The evolution of the music faculty: a comparative perspective," *Nature neurosciences*, 6, 663–668, 2003

[7] Acquisition of Communication and Recognition Skills Project (ACORNS) http://www.acorns-project.org/

[8] Human Speechome Project http://www.media.mit.edu/press/speechome/

[9] M. Kato, "Phonological development and its disorders," *J. Communication Disorders,* 20, 2, 84–85, 2003

[10] S. E. Shaywitz, *Overcoming dyslexia,* Random House, 2005

[11] M. Hayakawa, "Language acquisition and matherese," *Language*, 35, 9, 62–67, Taishukan pub., 2006

[12] P. Lieberman, "On the development of vowel production in young children," in *Child Phonology vol.1,* edited by G. H. Yeni-Komshian, J. F. Kavanagh, and C. A. Ferguson, Academic Press, 1980

[13] K. Okanoya, "Birdsongs and human language: common evolutionary mechanisms," *Proc. Spring Meet. Acoust. Soc. Jpn.,* 1-17-5, 1555–1556, 2008

[14] W. Gruhn, "The audio-vocal system in sound perception and learning of language and music," *Proc. Int. Conf. on language and music as cognitive systems*, 2006

[15] K. Miyamoto, *Making voices and watching voices,* Morikawa Pub., 1995

[16] T. Grandin *et al.*, *Animals in translation: using the mysteries of autism to decode animal behavior,* Scribner, 2004

[17] S. Umesh *et al.*, "Scale transform in speech analysis," *IEEE Trans. Speech and Audio Processing*, 7, 1, 40–45, 1999

[18] T. Irino *et al.*, "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: the stabilised wavelet-Mellin transform," *Speech Communication*, 36, 181–203, 2002

[19] A. Mertins *et al.* "Vocal trace length invariant features for automatic speech recognition," *Proc. ASRU*, 308–312, 2005

[20] R. Jakobson *et al.*, *The sound shape of language,* Mouton De Gruyter, 1987

[21] P. Ladefoged *et al.*, "Information conveyed by vowels," *J Acoust. Soc. Am.* 29, 1, 98–104, 1957

[22] J. Hawkins *et al.*, *On intelligence,* Henry Holt, 2004

[23] Y. Qiao *et al.*, "f-divergence is a generalized invariant measure between distributions," *Proc. INTERSPEECH*, 1349–1352, 2008

[24] Y. Qiao *et al.*, "A study of Hidden Structure Model and its application to labeling sequences," *Proc. ASRU*, 2009 (submitted)

[25] S. Asakawa *et al.*, "Multi-stream parameterization for structural speech recognition," *Proc. ICASSP*, 4097–4100, 2008

[26] N. Minematsu *et al.*, "Implementation of robust speech recognition by simulating infants' speech perception based on the invariant sound shape embedded in utterances," *Proc. SPECOM*, 35–40, 2009

[27] *Tohoku university – Matsushita isolated Word database (TMW)*, http://research.nii.ac.jp/src/eng/list/detail.html#TMW

[28] D. Ropar *et al.*, "Do individuals with autism and Asperger's syndrome utilize prior knowledge when pairing stimuli?" *Developmental Science*, 4, 4, 433–441, 2001

[29] U. Frith, *Autism: explaining the enigma,* Blackwell Pub., 1992

[30] N. Higashida *et al.*, *Messages to all my colleagues living on the planet*, Escor Pub., 2005

[31] L. H. Willey, *Pretending to be normal: living with Asperger's syndrome,* Jessica Kingsley Pub., 1999

[32] O. Sacks, *An anthropologist on mars,* Vintage, 1996

[33] S. Kitazawa, "Psychology and neuroscience to support autistics," Trans-disciplinary symposium on brain science and society, 2008