Are Learners Myna Birds to the Averaged Distributions of Native Speakers? — A Note of Warning from a Serious Speech Engineer —

Nobuaki Minematsu

Graduate School of Frontier Sciences, The University of Tokyo

mine@k.u-tokyo.ac.jp

Abstract

The current speech recognition technology consists of clearly separate modules of acoustic models, language models, a pronunciation dictionary, and a decoder. CALL systems often use the acoustic matching module to compare a learner's utterance to the templates stored in the systems. The acoustic template of a phrase is usually calculated by collecting utterances of that phrase spoken by native speakers and estimating their averaged distribution. If phoneme-based comparison is required, phoneme-based templates should be prepared and Hidden Markov Models are often adopted for training the templates. In this framework, a learner's utterance is acoustically and directly compared to the averaged distributions. And then, the notorious mismatch problem more or less inevitably happens. I wonder whether this framework is pedagogically-sound enough. No children acquire language through imitating their parents' voices acoustically. Male learners don't have to produce female voices even when a female teacher asks them to repeat her. What in a learner's utterance should be acoustically matched with what in a teacher's utterance? I consider that the current speech technology does not have any good answers and this paper proposes a good candidate answer by regarding speech as music.

1. Introduction

Many speech sounds are produced as standing waves in a vocal tract and their acoustic properties depend on the shape of the vocal tube. No two speakers have the same tube and therefore, speech acoustics vary among them. A process of producing a vowel sound is similar to that of producing a sound with a wind instrument. A vocal tube is an instrument and, by changing its shape dynamically, /aiueo/ is generated, for example. Different shapes cause different resonance, which causes different timbre. Acoustic differences in speakers are due to differences of the shape of the tube. Those in vowels of a single speaker are also for the same reason.

The aim of speech recognition is to extract only the linguistic information from speech. As speech contains both linguistic and extra-linguistic features, the current technology tries to extract only the linguistic information based on the following strategy,

$$g(\text{linguistic}) = \sum_{\text{extra-linguistic}} f(\text{linguistic}, \text{extra-linguistic}).$$

This is called *collectionism* and HMMs are a typical example. IBM ViaVoice collected speech samples from 350 thousands of American speakers. Many CALL products adopted ViaVoice as speech recognition engine and the above number is used even in advertisement [1]. As far as I know, however, no children acquire the ability to recognize speech after hearing 350 thousands of speakers. A major part of speech an infant hears is from its father and mother. After the infant begins to talk, as the speech chain implies, about a half of speech it hears will be its own speech. It is completely impossible for a human hearer to experience a speaker-balanced speech corpus. But the collectionism needs that for machines.

Why is a large corpus covering an enormous number of speakers needed? This is because the current speech technology does not have a good way to remove the speaker information from speech. Pitch information can be removed effectively by smoothing a given spectrum slice. Similarly, is there any good method to remove the extra-linguistic information from speech? What I'm discussing is not normalization or adaptation with respect to speakers. Spectrum smoothing is not a technique for normalizing pitch but for removing pitch. Given a smoothed spectrum, it is difficult to guess the pitch information included in the original speech. Is there any speech representation where it is difficult to guess who generated the speech sample? If one hears speech sounds, he can guess who produced them. This means that the desired representation may not include any factors which can reconstruct the sound substances but indicate only the linguistic skeleton of spoken language.

Developmental psychology tells that infants acquire spoken language through imitating the speech from their parents, called vocal imitation [2]. But no infants try to imitate the voices. As they have little phonemic awareness [3], they cannot identify a sound as phoneme although they can discriminate two different sounds. Namely, they cannot decode the speech into sequence of phonemes or convert the phonemes into sounds. In this situation, what in a father's speech is acoustically imitated by infants? Some researchers claim that they firstly learn the holistic sound pattern of the word [2], called word Gestalt. Then, what is the acoustic definition of that word Gestalt? If it includes speaker information, many infants must try to produce their fathers' voices. This consideration indicates that the word Gestalt has to be speaker-invariant. But what is that acoustically? I asked this question to many researchers in some conferences on infant study [4] but no researchers gave me a definite answer. If the word Gestalt could be defined acoustically, I'm wondering whether it might be the linguistic skeleton.

No infants imitate the voices but myna birds imitate not only the voices but also many sounds of cars, doors, animals, etc. Hearing a good myna bird say something, one can guess its keeper [5]. Hearing a very good child say something, however, it is impossible to guess its keeper. If one trains a myna bird to be a better imitator, the bird's voice and the target sound will be acoustically and directly compared and, to reduce the difference, some other training will be done. Most of the CALL systems directly compare an input utterance to the averaged distributions of many native speakers. This fact simply claims that the systems assume that a learner is a myna bird to the averaged distributions of the native speakers. Is this assumption correct and pedagogically-sound enough?

The problem I'm addressing is one of the fundamental but unsolved questions in speech science, which is variability of speech acoustics and invariance of speech perception [6]. I consider that, as this problem still remains to be unsolved, all the technical discussions have to be based on the collectionism. In the following sections, I propose a novel framework which can solve this problem by considering some similarities between language and music.

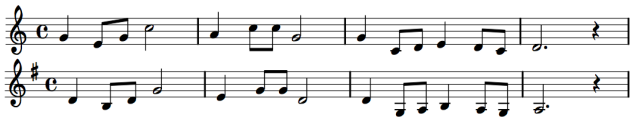
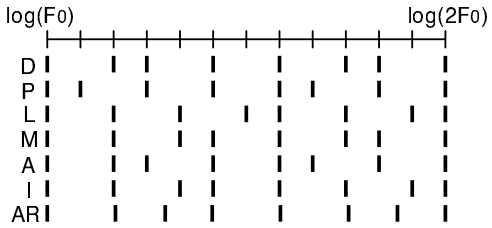


Figure 1: A musical piece and its transposed version



D=Dorian, P=Phrygian, L=Lydian, M=Mixolydian
A=Aeolian, I=Ionian, AR=Arabic

Figure 2: Various musical scales of 8 tones in an octave

Some people identify the individual tones in a given melody as syllable name. This identification is done independently of the key of that melody. Completely different tones are identified as Do, for example. In this paper, the mechanism of this key-invariant identification is considered for speech perception and recognition.

2. Relative sense of musical sounds

2.1. Key-invariant and robust identification of sounds

If one asks a number of people to transcribe the musical pieces in Figure 1 as sequence of Do, Re and Mi, what kind of answers are expected? Three kinds of answers are possible. Some will answer that the first one is So-Mi-So-Do La-Do-Do-So and that the second one is Re-Si-Re-So Mi-So-So-Re. These people are considered to have absolute pitch (AP) and Do, Re, and Mi are pitch names for them, i.e., *fixed Do*. Some others will claim that, for both pieces, they hear in mind the internal voices of So-Mi-So-Do La-Do-Do-So. It is interesting that they hear exactly the same sequence for physically different acoustic stimuli. They are considered 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 others will say "I cannot. I can only sing that sequence using La-La-La-La La-La-La-La." They also have RP but they cannot verbalize a musical piece or identify a sound as one of the sound categories.

AP people can memorize the absolute height of tones and use them to name musical sounds. RP people capture the difference of the height between two tones (musical interval). If one explicitly defines the acoustic height of the Do sound, all the RP people, including "La-La" people, can verbalize a given melody by following that definition. The difference between the second and the third groups is that the former do not need any helper who defines Do acoustically. Why and how can they name the individual sounds without memories of the absolute height of tones? The answer owes to the tonic scale embedded in music and, because this scale structure is key-invariant, the second group of people can perform the key-invariant and robust identification of the incoming sounds.

Figure 2 shows seven musical scales, all of which consist of eight tones in an octave. The first six scales, D to I, are the medieval church scales and Ionian and Aeolian are known as major and minor scales in the modern music. In these scales, an octave is divided into 12 semitone intervals and 8 tones are arranged so that they have 5 wholetone intervals and 2 semitone ones. These scales can be played by a normal piano but the last scale, Arabic, cannot be played by a normal piano because, as shown in Figure 2, it requires some tones not corresponding to any piano keys.

The second group, who transcribe both of the musical pieces in Figure 1 commonly as So-Mi-So-Do La-Do-Do-So, keep the major and minor sound arrangements in remembrance and, based on these arrangements, they identify the individual sounds [7]. This is why they cannot identify a separate sound as syllable name. When only a few sounds are presented, they also perform unstable identification of the sounds [7]. Another interesting phenomenon is that they find it difficult to transcribe a musical stream for some time immediately after modulation in key. Even in this case, the people with AP can naturally transcribe the individual sounds as pitch names. On the contrary, they sometimes don't notice the key change. But their identification is key-dependent, not robust at all.

Musicologically speaking, the syllable names of Do, Re,..., Si are the nicknames of musical functions. Do is for Tonic sound, which is the representative sound of that melody. The other functions are defined mainly in relation to the tonic sound. For example, Fa and So are for Subdominant and Dominant. How are the interrelations among sounds represented acoustically and mathematically? The interrelation of two sounds are basically defined by their harmony, the degree of oneness a hearer perceives when the two sounds are generated simultaneously. Since the oneness is a perceptual image, it is very difficult to derive its mathematical interpretation. But many researchers consider that pitch ratio is an important factor. A tonic sound and another tonic sound one octave above have pitch ratio of 1:2. Similarly, Tonic and Dominant have 2:3 and Tonic and Subdominant have 3:4. These pitch ratios can be converted into pitch difference or contrast in the logarithmic scale. Figure 2 is shown in the logarithmic scale of pitch (F_0).

The key-invariant and robust identification of sounds is possible because the RP people have no absolute templates of the individual sounds and they dynamically capture the embedded key-invaraint scale structure based on the pitch contrast to perceive the musical functions of the incoming sounds [7]. Brain sciences claim that RP is possessed only by humans [8]. The other primates can hardly perceive the equivalence between the two musical pieces in Figure 1. "La-La" singing is impossible for them.

2.2. Musical implementation of robust speech processing

In speech science and engineering, as far as I know, all the discussions of the speech sound identification were done based on the absolute identification. Is it possible to discuss the relative identification of speech sounds based on the speaker-invariant sound structure with no templates of separate sounds? Music is dynamic changes of pitch. Similarly, speech is dynamic changes of timbre. In Figure 3, a piano sound sequence of CDEFG and a speech sound sequence of /aiueo/ are shown, respectively. Dynamic changes are visualized in a phase space. Here, pitch is a one-dimensional feature of F_0 and timbre is tentatively shown as two-dimensional feature of F_1 and F_2. Cepstrum coefficients can also be used to expand a 10- to 20-dimensional phase space. Transposition of music translates the dynamic changes of F_0 and the shape of the dynamics is not altered. If the extra-linguistic factors of speech cannot change the shape of the speech dynamics, the relative identification based on the invariant sound structure can be used to implement super-robust speech processing. In the following section, a mathematical framework for the relative processing is described, where the robust invariance with respect to the extra-linguistic factors is guaranteed mathematically. After that, some experimental discussion evaluates the validity of the proposed framework.

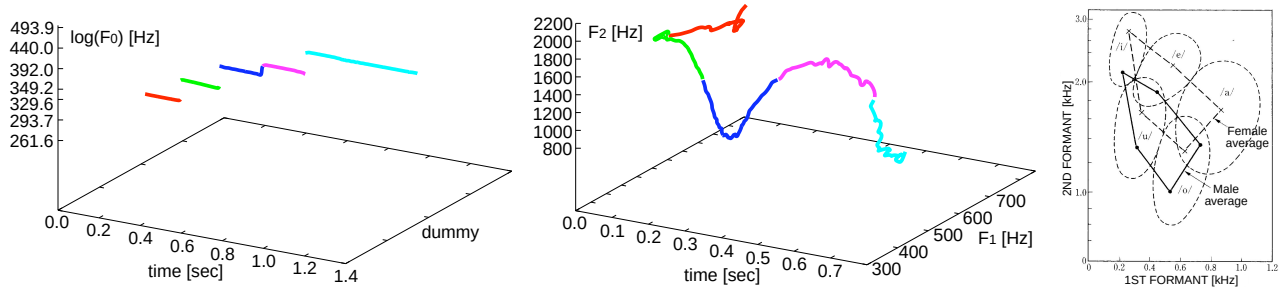


Figure 3: Dynamic changes of pitch in CDEFG and those of timbre in /aiueo/ with the Japanese vowel chart

3. Robust and structural invariance

3.1. Mathematical derivation of the invariant structure

In the above discussion, the transposition of music is supposed to correspond to the change of speaker, microphone, etc. In the case of music, the robust invariance of the pitch dynamics is given. The question is whether it can be obtained with speech. As shown in the Japanese vowel chart in Figure 3, it is often said in phonetics that the male vowel structure can be translated to become the female vowel structure. If this is correct enough, the timbre dynamics can be easily formulated to be invariant because speaker difference only translates the sound structure, namely, multidimensional transposition. But every speech engineer knows that this idea is so simple that it cannot be applied effectively to real speech data.

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 the speaker conversion research in speech synthesis. Figure 4 shows two acoustic spaces of speakers A and B. Acoustic events of p_1 and p_2 of A are mapped to those of q_1 and q_2 of B, respectively. It is easily supposed that a mapping function of A's entire space into B's entire space has to be very complicated and strongly dependent on both the speakers. This indicates that, if one wants to focus on the invariance of the timbre dynamics, he has to derive some invariant acoustic observations with respect to any form of mapping function. Is the *robust* invariance possible?

The answer is yes if the two spaces have one-to-one correspondence [9]. (x, y) in space A is uniquely mapped to (u, v) in space B and vice versa. Every event is characterized as distribution.

$$1.0 = \iint p_i(x,y)dxdy, \quad 1.0 = \iint q_i(u,v)dudv$$

Here, we consider functions of f and g for the mapping, i.e. $x{=}f(u,v)$ and $y{=}g(u,v)$. f and g can be non-linear. Even when they cannot be represented by any known analytical expressions, the following discussion is effective. Using f and g, any integral operation in space A can be rewritten as its counterpart in space B.

$$\begin{aligned}\iint \phi(x,y)dxdy &= \iint \phi(f(u,v),g(u,v))|J(u,v)|dudv \\ &= \iint \psi(u,v)dudv,\end{aligned}$$

where $\psi(u,v) = \phi(f(u,v),g(u,v))|J(u,v)|$. $J(u,v)$ is Jacobian. Any p_i in A can be mapped into q_i in B.

$$q_i(u,v) = p_i(f(u,v),g(u,v))|J(u,v)|.$$

Physical properties of p_i are different from those of q_i. For example, p_1 may represent /a/ of speaker A and q_1 may represent /a/ of

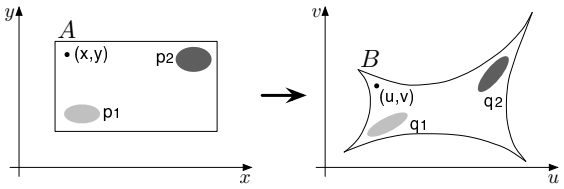


Figure 4: Linear or non-linear mapping between two spaces

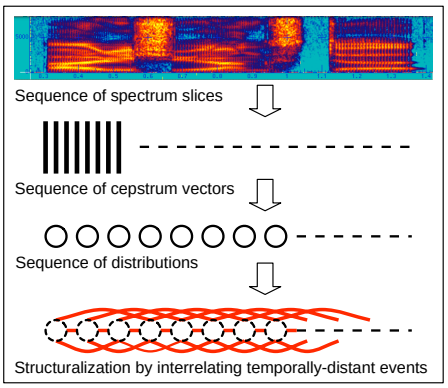


Figure 5: BD-based robustly-invariant structure of speech

speaker B. What can be robustly invariant between a set of p_is in space A and a set of q_is in space B? Let us consider Bhattacharyya distance between two events (distributions).

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

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

$$= -\log \iint \sqrt{q_1(u,v)q_2(u,v)}dudv = BD(q_1,q_2)$$

BD between two events in space A and BD between their corresponding two events in space B cannot be changed. Events can change easily but their difference or contrast cannot change by any transformation. The shape of a triangle is determined uniquely if the length of the three segments is given. Similarly, the shape of an n point geometrical structure is determined uniquely if the length of all the ${}_nC_2$ segments, including the diagonal ones, is given. In other words, if a distance matrix is given for n points, the matrix determines the shape of the n-point structure. As told above, BD is robustly transformation-invariant. Given n distributions, a BD-based distance matrix represents its robustly-invariant structure.

3.2. Experimental verification of the structure

From a spoken utterance, it is possible to extract its invariant structure, shown in Figure 5. After converting the utterance into a se-

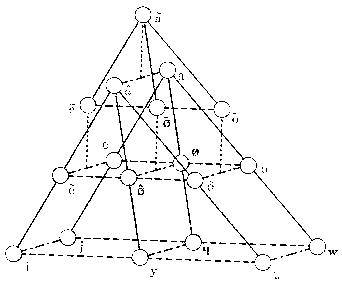


Figure 6:Jakobson's geometrical structure of the French vowels [10]

quence of distributions, all the timbre contrasts between any two distributions are calculated to form an invariant speech structure (distance matrix). Here, long-distance contrasts are also considered here. The mathematical fact that any transformation cannot change the structure indicates that any transformation works geometrically as either of the two operations, rotation and shift. Using this geometrical property, the structural matching between a word utterance and another was implemented and tested by recognizing utterances of 5 connected Japanese vowels. It was surprising that 99.3% of the vowels were correctly recognized without the use of any absolute properties of the sound substances [9]. The number of training speakers was 8 and the performance was better than that of HMMs trained by 4,130 speakers (98.8%). This machine cannot identify an isolated sound at all because it has no structure.

Many people can identify the incoming musical sounds only based on the pitch contrasts and their invariant structure. Similarly, the above machine can identify the incoming speech sounds based on the timbre contrasts and their invariant structure. Phonetics discusses the absolute values of the linguistic sounds and phonology does their interrelational values. The conventional speech engineering is based on phonetics and the proposed framework is based on phonology. For example, Figure 6 shows Jakobson's structure of the French vowels, i.e, his skeleton of the linguistic sounds.

3.3. Other interesting discussions

Some researchers of brain sciences claim that, on the cortex, the linguistic aspect and the extra-linguistic aspect of speech are separated and the former is encoded as motions in speech [11]. If their claim is valid enough, the ability of identifying an isolated sound as phoneme is not needed for language competence. Speech communication without that ability was experimentally verified [12, 13]. Technically speaking, speech samples of very large people like giants and very small people like fairies can be easily generated. It is interesting that the isolated vowels produced by these people could not be correctly identified by listeners. With 65 [cm] people, the identification rate was chance level because the range of F_1/F_2 was by far out of the range of real people. Once they uttered even a *meaningless* sequence of sounds continuously, however, the sound identification rate drastically improved. This result indicates that human speech *stream* perception is not a process of sequential and separate identification of the incoming sounds.

Figure 2 shows various patterns of the musical scale. Within an octave range, the sound arrangement differs among them. If Western music is played with the Arabic scale, it will be called Arabic accented Western music. A similar discussion is possible with vowels. The variation of the F_1/F_2 vowel structure of a language represents the variation of its regional accents. English with the Arabic vowel structure is called Arabic accented English.

4. Application to the CALL research

The new framework has been already applied to CALL because it was originally proposed for CALL. The system with acoustic models trained with speech substances, in principle, has to have the mismatch problem. Considering that the syllable name identification is performed only with the key-invariant pitch contrasts without any normalization or adaptation, I implemented a similar framework for speech and verified it experimentally. After that, I found that what I proposed was phonology-based speech engineering. Some CALL papers were already presented in major speech conferences and the interested readers should refer to [14, 15, 16], for example. In these works, the pronunciation portfolio was proposed, where the pronunciation development is logged, the adequate instructions on what to do next are generated, and the classification of the learners is done irrespective of speaker differences.

The people with extremely AP have difficulty in perceiving the equivalence between the two musical pieces in Figure 1. Similarly, those with extremely absolute sense of speech sounds have difficulty in perceiving the equivalence between their mothers' "Good morning" and their fathers' "Good morning". Some autistics have that difficulty. An autistic boy *wrote* that it is easy to recognize his mother's speech but difficult to do his father's [17]. But it is also difficult to recognize his mother's speech on telephone line. Autistics have very good absolute sense of sounds and are often very good at copying sounds [18]. [19] describes another autistic boy who imitates not speech but voices like myna birds. In most of these cases, they don't have spoken language. I have to wonder whether the conventional CALL systems assume that the learners are autistic to the averaged distributions of the native speakers.

5. References

[1] http://tepia.or.jp/archive/12th/pdf/viavoice.pdf
[2] P. W. Jusczyk, *The discovery of spoken language*, Bradford Books (1997)
[3] S. E. Shaywitz, *Overcoming dyslexia*, Random House Inc. (2005)
[4] N. Minematsu *et al.*, "Universal and invariant representation of speech," Proc. Int. Conf. Infant Study (2006)
http://www.gavo.t.u-tokyo.ac.jp/˜mine/paper/PDF/2006/ICIS_t2006-6.pdf
[5] K. Miyamoto, *Making voices and watching voices,* Morikita Pub. (1995)
[6] K. Johnson and J. W. Mullennix, *Talker variability in speech processing*, Academic Press (1997)
[7] T. Taniguchi, *Sounds become music in mind – introduction to music psychology –*, Kitaoji Pub. (2000)
[8] D. J. Levitin *et al.*, "Absolute pitch: perception, coding, and controversies," Trends in Cognitive Sciences, 9, 1, pp.26–33 (2005)
[9] S. Asakawa *et al.*, "Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics," Proc. InterSpeech (2007, to appear)
[10] R. Jakobson *et al.*, *Notes on the French phonemic pattern*, Hunter (1949)
[11] P. Belin, *et al.*, "'What', 'where' and 'how' in auditory cortex," Nature neuroscience, 3, 10, pp.965–966 (2000)
[12] D. Smith *et al.*, "The processing and perception of size information in speech sounds," J. Acoust. Soc. Am., 117(1), pp.305–318 (2005)
[13] Y. Hayashi *et al.*, "Comparison of perceptual characteristics of scaled vowels and words," Proc. Spring Meeting Acoust. Soc. Jpn., pp.473–474 (2007)
[14] N. Minematsu, Proc. ICSLP, pp.1669–1672, pp.1317–1320 (2004)
[15] S. Asakawa *et al.*, Proc. EuroSpeech, pp.165–168 (2005)
[16] N. Minematsu *et al.*, Proc. IEEE Int. Workshop on Spoken Language Technology, pp.126–129 (2006)
[17] N. Higashida *et al.*, *Messages to all my colleagues living on the planet*, Escor Pub. (2005)
[18] U. Frith, *Autism: explaining the enigma*, Blackwell Pub. (1992)
[19] T. Asami, *A book on my son, Hiroshi*, vol.5, Nakagawa Pub. (2006)