CONSTRUCTING SPEECH BANANA FOR THAI CONSONANTS: SOME CONSIDERATIONS FOR MALE AND FEMALE VOICES

N. Klangpornkun[1], C. Onsuwan[2, 3], C. Tantibundhit[1,3]

[1] Department of Electrical and Computer Engineering, Thammasat University, Thailand
[2] Department of Linguistics, Thammasat University, Thailand
[3] Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Thammasat University, Thailand
consuwan@tu.ac.th, tchartur@engr.tu.ac.th

ABSTRACT

"Speech banana" is a banana-shaped plot of speech power distribution. It shows speech sounds spoken with normal loudness in terms of frequency and intensity level. Speech bananas have been mainly proposed for Indo-European languages. This work describes an ongoing work of constructing a speech banana for Thai, a Non-Indo-European language. Speech materials are taken from 11 Thais (males and females). Comparisons are carried out to examine if there is any difference in speech power distribution of 21 consonants from male and female voices. The findings show statistically significant differences in terms of level of intensity for 13 phonemes (/m/, /b/, /n/, /d/, /ŋ/, /j/, /r/, /w/, /l/, /k/, /h/, /pʰ/, /f/) and in terms of frequency for /s/ only. It is suggested that significant acoustic differences associated with male and female voices should be incorporated in representation of speech banana.

Keywords: speech banana, Thai, power spectral density, intensity, frequency, male and female

1. INTRODUCTION

1.1. Background

"Speech banana" is a banana-shaped plot of speech power distribution, where the abscissa and ordinate represent frequency (Hz) and intensity (dB). In other words, it shows speech sounds (consonants and vowels) spoken with normal loudness in terms of frequency and intensity level [1, 2, 3, 4]. Although in reality, speech is a stream of sounds, with one sound following another in a rapid fashion, each speech sound seems to occupy certain locations on the speech banana (i.e., vowels tend to have a lower frequency and a higher intensity level than consonants) [5].

Despite the fact that speech banana has been widely accepted and referred to in the fields of audiology and hearing sciences [3, 4, 5], techniques and steps that were employed in constructing each speech banana are not well documented [4]. To our knowledge, many versions of speech banana exist for English, one of which was proposed by Northern and Downs [3]. The best-known speech banana is for Swedish sounds by Liden and Fant [1]. However, there appears to be no study that describes speech banana for Non-Indo-European languages or for tonal languages, such as Chinese and Thai.

As languages across the world are known to have unique phonemic systems, it should not be surprising that each language's speech banana will differ to some extent from one another. It is reasonable to expect that the (overall) figurative areas might be similar, but specific locations for speech sounds may vary.

1.2. Thai speech banana

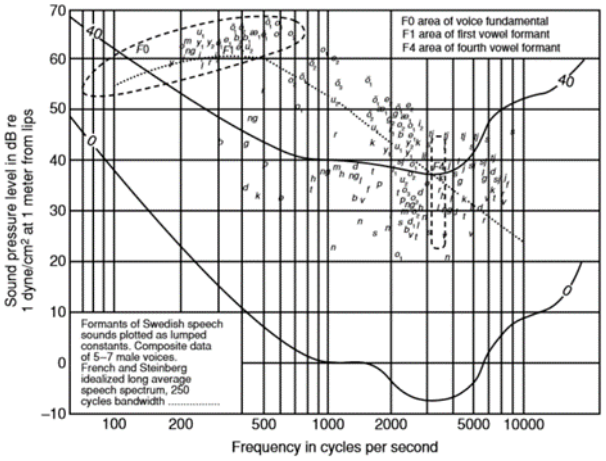
Recently in Klangpornkun *et al.*, we proposed a Thai speech banana specifying 21 consonants, 18 vowels, and 5 lexical tones [6]. In fact, /f/ /h/ and /s/ were among the softest sounds and vowels /ɛ/ /o/ and /ɔ/ among the loudest. It should be noted that our proposed speech banana was drawn from one male speaker.

1.3. Male and female voice

A large body of research has clearly shown differences between the anatomy, physiology, and acoustics of male and female voices [7]. Particularly, male vocal folds tend to be longer than females; therefore, female's pitch is generally higher. Some researchers reported that female voices show greater levels of aspiration noise [7, 8]. Moreover, (with the use of Long-Term average Spectrum) male and female voices showed significant differences in the distribution of energy for a long-term basis [7].

For the above mentioned reasons, this present study examines if there is any significant difference in speech power distribution of 21 consonants from male and female voices and proposes a way to represent the difference (if any) in a speech banana.

Figure 1: Sound pressure level versus frequency of Swedish vowel and consonant formant data (from Lidén and Fant [1]).



2. PREVIOUS SPEECH BANANA

Very few details regarding the construction of speech banana were discussed in the literature. In Lidén and Fant's work, the plot of sound pressure level (dB SPL) and frequency (Hz) of Swedish speech sounds was given (Fig. 1) and the data was used to construct the Swedish speech banana came from male talkers only [1].

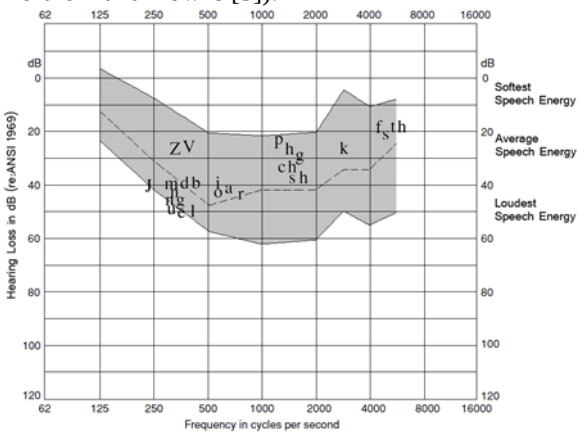
The plot consists of six parts: (1) speech formants from all Swedish vowels and consonants (2) average speech spectrum from French and Steinberg [9] (3) area of fundamental frequency of voiced phonemes (4) area of fourth vowel formant (5) a 0 phon equal loudness contour or standardized free filed threshold of hearing (6) a 40-phons equal loudness contour [1]. It is noteworthy that the equal loudness contour is the minimum threshold of hearing determined by the responses of several healthy young people, where the subjects sat facing to the source in a free field and judged that two sounds were equally loud or the sound was just audible [3]. The assigned sound pressure levels were the free-field levels unaffected by diffraction effects due to the presence of the auditor [3].

Northern and Downs proposed a construction of English speech banana (Fig. 2) [3] which was adapted from Dudich *et al.,* and Skinner [10] (with no mentioning of talker(s)' gender). In addition, they showed the average range between softest and loudest energy of conversational speech.

2.1. Speech material in previous study

As mentioned, very little is known regarding the construction of previous speech bananas. To our knowledge, the Swedish speech banana was constructed based on 5-7 male voices recorded at 1 meter from lips [1]. Therefore, in the present study,

Figure 2: Average range of (English) speech energy with the average range of softest and loudest speech energy in conversation (from Northern and Downs [3]).



we believe that it is important to incorporate data from a larger number of speakers and to include both male and female voices.

2.2. Thai phonology

Thai is a tonal language with 21 consonantal phonemes in initial position /p/, /pʰ/, /b/, /t/, /tʰ/, /d/, /tɕ/, /tɕʰ/, /k/, /kʰ/, /ʔ/, /f/, /s/, /h/, /m/, /n/, /ŋ/, /l/, /r/, /w/, and /j/. Out of 21 consonants, nine occur in final position. Each of the nine monophthongs in Thai occurs phonemically short or long. There are five lexical tones: mid, low, high, falling, and rising. Thus, Thai syllables maybe represented as $C_i(C)V^TC_f$ or $C_i(C)V{:}^T$,where C_i stands for an initial consonant, C_iC aconsonantal cluster, C_f a final consonant, V a short vowel, V: a long vowel, and T a tone.

3. THAI SPEECH BANANA (TSB) CONSTRUCTION

3.1. Speech material

We construct a TSB of 21 initial consonants from recordings of six Thai males (one speaker also participated in the study by Klangpornkun *et al.* [6]) and five Thai females. All speakers were born and grew up in Bangkok. The average age is 22.18 years old with 4.62 years SD. Speech materials were recorded at a sampling rate of 44.1 kHz in a sound-attenuated chamber at a conversational level (around 60–80 dB SPL). There are 21 target words of the form [Cāː] differing only in their initial consonants (e.g., /bāː/, /fāː/). The words were embedded in a carrier sentence [tɕʰɔ̂ːp ... ʔìːik lɛ́ːw] to assure that it sounds as natural as possible [11]. Each speaker uttered each of 21 sentences five times. The target

Figure 3: Box plot of intensity level in dB SPL of 21 initial rhyming words (CV:) from 11 speakers prior to (top) and after (bottom) adjusting the intensity level.

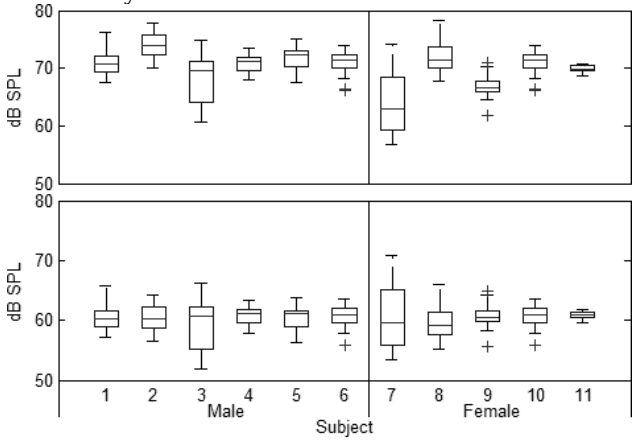
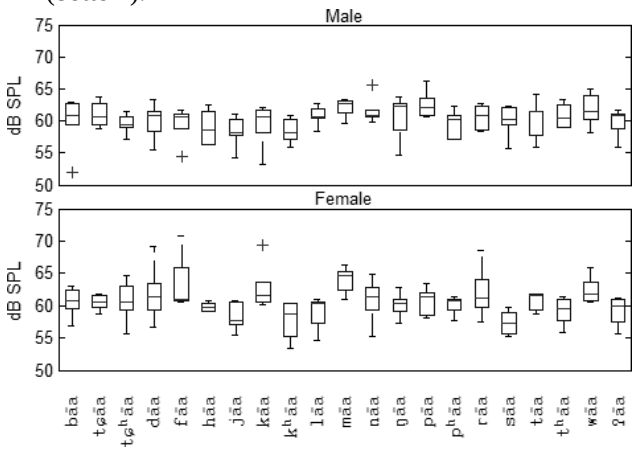


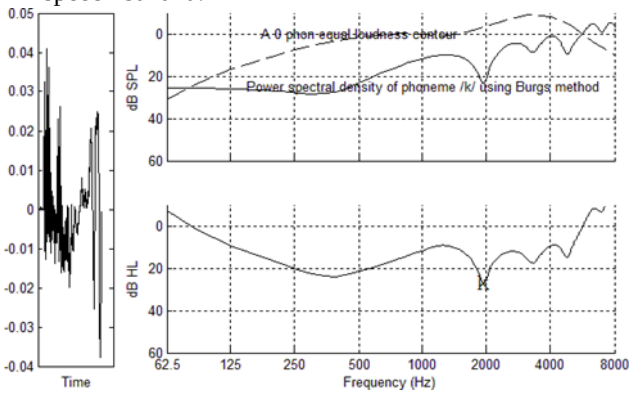
Figure 4: Box plot of intensity level in dB SPL of 21 target words from males (top) and females (bottom).



words were manually excised from their carrier sentences. One repetition of each word (the best token) was chosen based on impressionistic hearing evaluation and spectrographic inspection [11] resulting in 21 target words per speaker. Then, the average intensity of all words was adjusted to 60 dB SPL as according to the normal conversation level [3]. A note should be made regarding some allophonic variation of /r/ among Thais, which is also observed in our recordings. Our male takers produce /r/ as a flap or tap whereas female talkers a trill.

Figure 3 illustrates box plot of intensity level in dB SPL across 21 target words of each speaker before and after the intensity adjustment to 60 dB SPL. Figure 4 illustrates box plot of intensity level in dB SPL across 21 target words from males (top) and females (bottom). It should be noted that four speech tokens are outliers (each from /bāː/, /fāː/ and /nāː/ by male speakers and one token of /kāː/ by a female speaker) and they were removed and will not be used in constructing TSB. Finally,

Figure 5: (a) Time domain of phoneme /k/, (b; top) power spectral density (solid line) and the intensity of a 0 phon equal loudness contour (dotted line), (b; bottom) difference between the intensity of power spectral density (solid line) and position of phoneme /k/ at the local maxima in speech banana.



speech portions of the initial consonants were manually excised from the selected words.

3.2. Acoustic analysis

In Klangpornkun *et al*, frequency and intensity (dB SPL) of each phoneme were referred to as the local maximum of its power spectral density [6] and then were located one by one to give a speech banana. In this present study, we refine the procedure so that each point can be as precise as possible. Details of constructing consonants TSB are given bellows:

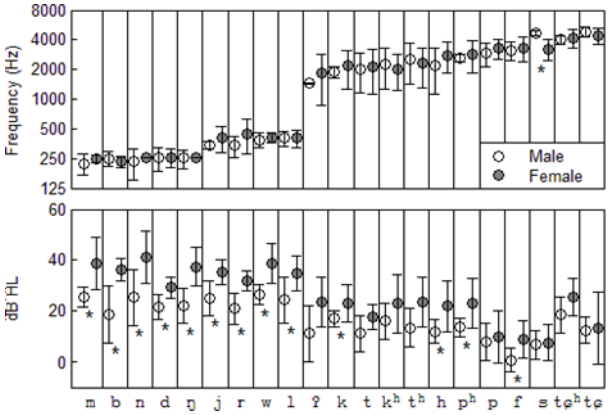
1. Calculate power spectral density of a given phoneme using Burgs method with order 40. This order was chosen based on [6].
2. Calculate the difference between power spectral density (dB SPL) of the phoneme and the intensity of a 0 phon equal loudness contour (dB SPL) as shown at the top of Fig. 5 (b).
3. Plot the resulting power spectral density (dB HL) of the phoneme as shown at the bottom of Fig. 5 (b).
4. Locate its local maximum of the power spectral density from Step 3.
5. Locate the phoneme position precisely at the local maximum on the plot.

4. RESULTS

4.1. Male and female voice

Two-sample t-test was performed (with 0.05 level of significance) on average frequency and average intensity (dB HL) for each initial phoneme in order to address the issue of whether the difference between male and female voices significantly affects the level of frequency and intensity. The results are

Figure 6: 95% CI of frequency (Hz) and intensity level (dB HL) represented as pairs of bars between male (left) and female (right) across 21 initial consonants.



shown in Figure 6. Only the frequency of /s/ shows a statistically significant difference [t(9) = 4.9024, p = 0.0008] with a higher frequency for female voices. Interestingly, 13 phonemes are significantly different in terms of intensity level, i.e., /m/ [t(9) = -3.51, p = .01], /b/ [t(8) = -4.05, p = .01], /n/ [t(8) = -2.98, p = .02], /d/ [t(8) = -2.99, p = .02], /ŋ/ [t(9) = -4.03, p = .003], /j/ [t(9) = -3.14, p = .01], /r/ [t(9) = -3.66, p = .01], /w/ [t(9) = -4.01, p = .003], /l/ [t(9) = -2.37, p = .04], /k/ [t(8) = -2.57, p = .03], /h/ [t(9) = -2.62, p = .03], /p^h/ [t(9) = -2.63, p = .03], and /f/ [t(8) = -2.56, p = .03], respectively, all showing a higher intensity level for female voices.

4.2. Thai speech banana (TSB)

Figures 7-9 illustrate TSBs constructed from six males, five females, and all speakers, respectively. In the figures, individual phoneme is represented as separate shaded area, where the width in x-axis shows a 95% CI of frequency (Hz) and the width in y-axis shows a 95% CI of intensity level (dB HL). Shading levels represent CIs; darker shade shows narrower CI. Envelope of the TSB includes 95% CIs of all phonemes in terms of frequency and intensity level. As expected from the results in 4.1, frequency-wise (horizontal dimension), locations of the 21 phonemes on these three figures are quite the same. The notable difference is in terms of intensity (vertical dimension). Female voices exhibit higher intensity level as the banana is shifted downward.

5. DISCUSSIONS

This work provides detailed description of a process involved in constructing a speech banana for Thai, currently focussing on consonants. As the speech materials are taken from males and females, we are able to examine acoustic differences associated with gender, which have been overlooked in previous studies. We found significant differences in intensity level for 13 phonemes and in terms of frequency for /s/ only. To a certain degree this is to be expected given differences between male and female anatomy and physiology. However, it is not obvious to us as to why the differences manifest greatly in terms intensity and minimally in frequency. More analysis will be carried out to explore this further along with an analysis of Thai vowels and tones to be included on our Thai speech banana.

Figure 7: Thai speech banana from male speakers.

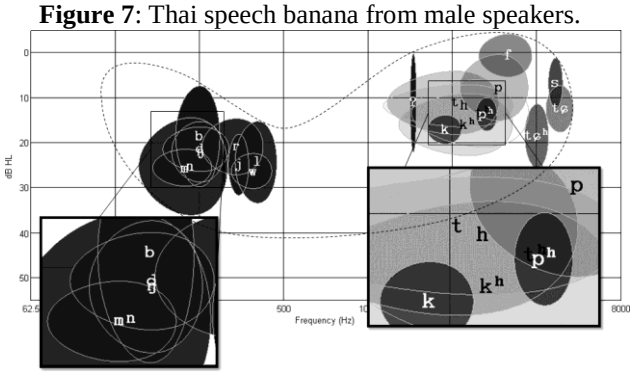


Figure 8: Thai speech banana from female speakers.

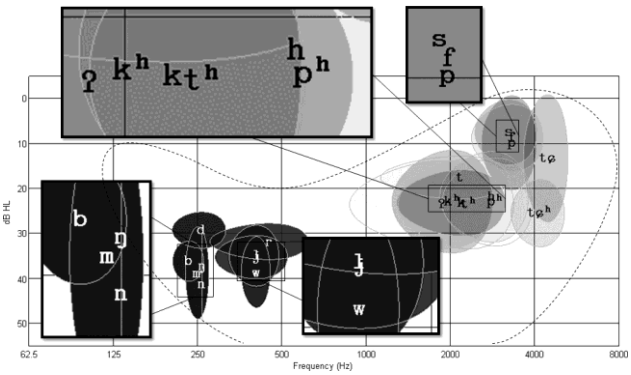
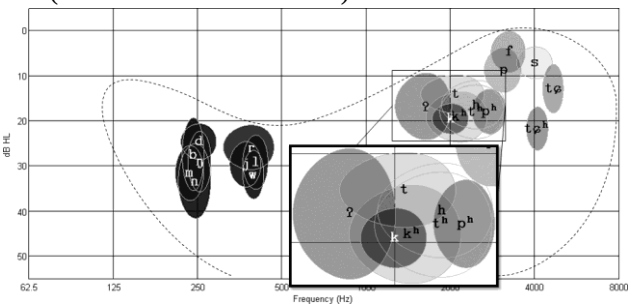


Figure 9: Thai speech banana from all speakers (six males and five females).



6. ACKNOWLEDGEMENT

We would like to acknowledge with much appreciation Chaipat Suwannaphum and Chinnawat Nualta (Department of Electrical and Computer Engineering, Thammasat University) for their role in data preparation process.

7. REFERENCES

[1] Lidén, G. and Fant, G. 1954. Swedish word material for speech audiometry and articulation tests. *Acta Oto-Laryngologica 43*. S116, 189–204.

[2] Ling, D. 1989. *Foundations of Spoken Language for Hearing-impaired Children* Alexander Graham Bell Association for the Deaf: Washington, DC.

[3] Northern, J. L. and Downs, M. P. 1984 *Hearing in Children* Lippincott Williams & Wilkins: Pennsylvania.

[4] Fant, G. 2004. Speech Perception. In: Fant G., *Speech Acoustics and Phonetics*.Kluwer Academic Publishers: Dordrecht, the Netherlands, 199–220.

[5] Ross, M. 2004. The Audiogram: Explanation and Significance. *Hearing Loss Association of America*. 25, 29–33.

[6] Klangpornkun, N., Onsuwan, C., Tantibundhit, C. and Pitathawatchai, P. 2014. Predictions from "speech banana" and audiograms: Assessment of hearing deficits in Thai hearing loss patients. *Proceedings of Meetings on Acoustics*. 20.

[7] Mendoza, E., Valencia, N., Muñoz, J. and Trujillo, H. 1996. Differences in Voice Quality between Men and Women: Use of the Long-Term Average Spectrum (LTAS). *Journal of Voice*. 10, 59–66.

[8] Klatt, D. H. and Klatt, L. C. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*. 87, 820–857.

[9] French, N. R. and Steinberg, J. C. 1947. Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*. 19, 90–119.

[10] Skinner, M. 1978. The hearing of speech during language acquisition. *Otolaryngol Clinics of North America*. 11, 631–650.

[11] Tantibundhit, C., Onsuwan, C., Saimai, T., Saimai, N., Thatphithakkul, S., Chootrakool, P., Kosawat, K. and Thatphithakkul, N. 2011. Perceptual Representation of Consonant Sounds in Thai. *Proc. INTERSPEECH* Florence, Italy.