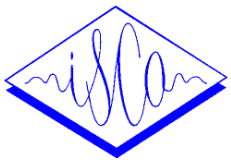

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

POLE-ZERO STRUCTURE BASED ON TWO-SOURCE VOCAL TRACT MODEL
—— PSE Inspection of Continuous Speech Vowel Part ——

Takayuki NAKAJIMA and Hiroshi OHMURA

Electrotechnical Laboratory
1-1-4 Umezono,Tsukuba Science City, 305, Japan

Abstract:

In this paper, at first, Japanese continuous speech vowel parts PSE (power spectrum envelope) are observed, and it is pointed out that there is a great difference between the continuous speech vowel PSE and the PSE of a vowel uttered in isolation. In order to explain the phenomena, a two-source vocal tract model is proposed. In the model, the vocal tract is expressed as a non-uniform single tube including the under glottal portion. Giving the vowel vocal tract area function corresponding to the 5 vowels of Japanese and two other parameters: the source position and the power ratio, vocal tract pole-zero transfer functions are calculated and an articulatory/acoustic nomogram is constructed. As the results of the inspection if there is a correspondence between the PSE of real speech and one of the nomogram's frequency transfer characteristics in pole-zeroes, the authors present the following hypothesis: "The continuous speech vowel part PSE has the kind of pole-zero structure generated by the two-source vocal tract model".

1 Introduction

Understanding a phoneme from the standpoint of its physical features is essential for realizing high level speech recognition and synthesis. The authors have proposed that phoneme understanding must be started based on a high precision PSE which reflects the pole-zero structure of the speech power spectrum generated from the human speech production system. For the object, the authors have presented a high precision pole-zero analysis named "Pitch pair synchronous PSE speech analysis method"[1]. In the method, at first stage, spectrum data series valid for the PSE estimation are extracted by pitch frequency interval sampling of the power spectrum which is obtained by applying a 2048 point FFT to the short term speech wave which is cut out by the Hamming window of which length is proportional to pitch interval; and cosine series as logarithmic PSE model is fitted to the above logarithmic spectrum data series at the second stage. Besides, the authors have developed a frequency transfer function calculation method of the vocal tract in an arbitrary source position by recursive equations[2].

In this paper, at first, Japanese continuous speech vowel part PSE are observed using the speech analysis method[1], and it is pointed out that there are great differences between the continuous speech vowel PSE and the PSE of a vowel uttered in isolation and also of past vowel model in pole-zero structure. In order to explain why the differences are created, a two-source vocal tract model is proposed.

2 Continuous Speech Vowel PSE Deviates from Past Model's

The speech data consists of Japanese news sentences: " *tada ima o tsutae shimashita you ni, atarashii douro koutsuu hou ga kyou kara jisshi sare mashita ga, keishi chou dewa shiro bai hyaku go juu dai wo douin shi.....*" read by two non-professional male adult speakers. In the above Roman character description, space " " means word boundary without acoustical separation.

As the results of visual inspection of the PSE data obtained by applying the pitch pair synchronous PSE analysis[1] to the above Japanese news sentences, the following phenomena are observed in the so-called vowel parts:

1) There are systematic zeroes growing from the so-called vowel part toward the following closed consonant. See **Fig. 1** , **Fig. 3**, and **Fig. 4**.

2) There systematic zeroes in the so-called vowel part which is preceded by a closed consonant. See **Fig. 2** and **Fig.5** a look.

3) There is a weakening or a lack of a specific pole. See **Fig. 3** a look.

4) An additional new pole is observed in the neighbourhood of the generated zero. See **Fig. 3**, and **Fig. 5**.

The major reason common to the above phenomena is estimated that generated zeroes give rise to several effects on a specific pole or a whole PSE, which gives rise to greatly different PSE from that of an isolated vowel of the same speaker. In speech research up to now, spectral zeroes based on the following reasons have been made clear:

1) Zeroes caused by vocal cords wave.
2) Zeroes caused by under glottal portion.
3) Zeroes caused by nasalization.

Nevertheless, the zeroes pointed out in Fig. 1 to Fig. 5 don't correspond to these reasons, but may be generated by the pseudo sound source which would be configured as the direct effect of the sound source generation mechanism of the neighbouring closed consonant.

The well known vowel model is based on a steady state. In the model, the vocal tract is limited from lips to glottis and a single source is at the glottal part. Vocal tract area function is assumed to be continuous and to have no extreme constriction compared to that of closed consonants. As a result, the frequency transmission characteristics of the vocal tract model is of an all-pole type. Even though an arbitrary vocal tract shape is given to the past vocal tract vowel model, zeroes and a specific pole weakening in the vowel part cannot to be produced. If vocal tract length is assumed to be 17 cm, which is the average value of male adult's, the average frequency interval between neighbouring poles is 1 kHz. Therefore, the overall logarithmic power spectrum in logarithmic frequency scale is roughly straight. The continuous speech vowel part PSE is far from the above model's.

3 A Hypothetical Explanation of PSE

3.1 An articulatory/acoustic nomogram

Using the recursive calculation method[2] of frequency transfer characteristics of a vocal tract with an arbitrary source position and in realistic acoustic conditions, an articulatory/power spectrum nomogram is constructed. Vowel vocal tract shapes are introduced from those of the ETL Vocal Tract Analog Speech Synthesis[3].

Figure 6 is a part of the nomogram. For each Japanese 5 vowel, shown in the top row, two sources are assumed: the first one, **k1**, is fixed at the glottis, the second one, **k2**, is placed a position corresponding to lips, tongue tip, or tongue body. To each vowel, the frequency transfer characteristics are shown in the column. In the figure, the parameter is the intensity ratio η of the second source to the first source.

The nomogram shows that the poles are configured by the whole vocal tract, therefore, they are independent from the source position. Besides, zeroes are configured by a partial vocal tract corresponding to rear part of the source, being dependent on the source positions. In general, when the second source is at the lips, the first zero is at a low frequency, and whole zeroes are nearly coincident to the whole poles, respectively. When the second source is at the tongue body, the first zero is at a very high frequency. The first, second, third, etc. zeroes are much influenced by vowel phonemic differences as well, so that one or some of the zeroes affect a pole or some poles. This creates a dominant zero, the specific pole weakening in frequency transmission characteristics or an over all logarithmic power spectrum bending in logarithmic frequency axis.

3.2 A hypothetical explanation of vowel PSE based on a two-source vocal tract model

Each pattern of the real continuous speech PSE is compared with each of the nomogram if the PSE pole-zero structure nearly corresponds to one of the power spectrums in the nomogram of Fig.6 by visual inspection. As the result of the inspection, the following correspondences are obtained:

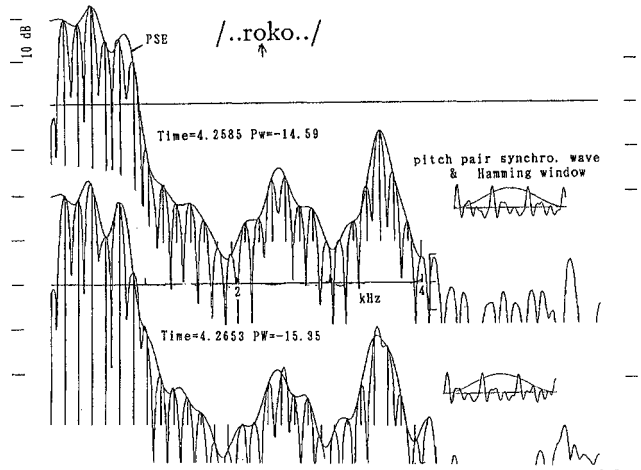

Fig. 1 Example of the pitch pair synch. PSE analysis[1]. Zeroes and their growth are observed in the vowel part in high power level. Its power level decreased to -31.81 dB at 4.2960 sec., -48.98 dB at 4.3135 owing to the following closed consonant /k/ using tongue body. (Corresponding nearly to the /u/, k2 = 5, η = 16., 276. of Fig. 6)

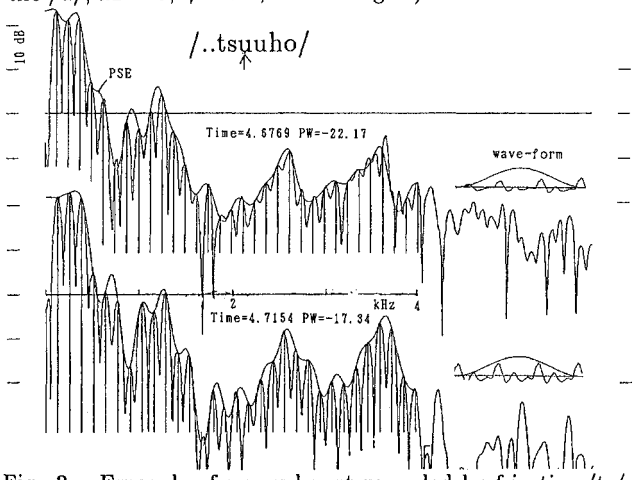

Fig. 2 Example of a vowel part preceded by fricative /ts/ by tongue-tip. Typical pole-zero structure is observed. The beginning of voicing is at 4.6696 sec. (Corresponding nearly to the /e/, k2 = 1, η = 8., 153. of Fig. 6)

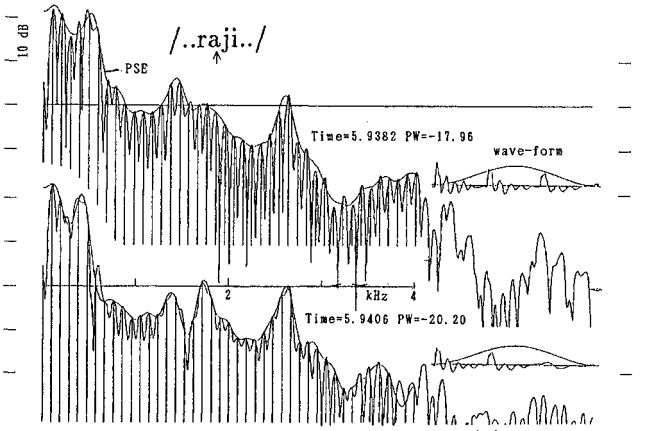

Fig. 3 Example of a vowel part followed by /j/ using tongue - tip; F_2 is canceled by zero. (Corresponding nearly to the /e/, k2 = 6, η = 146, 8 of Fig. 6)

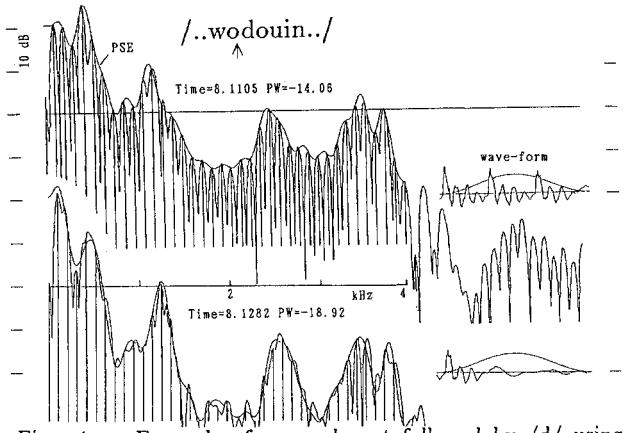

Fig. 4 Example of a vowel part followed by /d/ using tongue - tip; zeroes and their growth are observed. (Corresponding nearly to the /e/, k2 = 1, η = 8., 153. of Fig. 6)

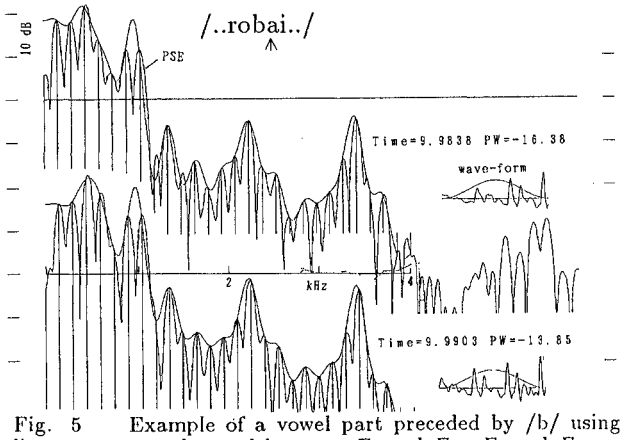

Fig. 5 Example of a vowel part preceded by /b/ using lips; zeroes are observed between F_1 and F_2 , F_2 and F_3, ... (Corresponding to the /a/, k2 = 0, η = 276., 16. of Fig. 6)

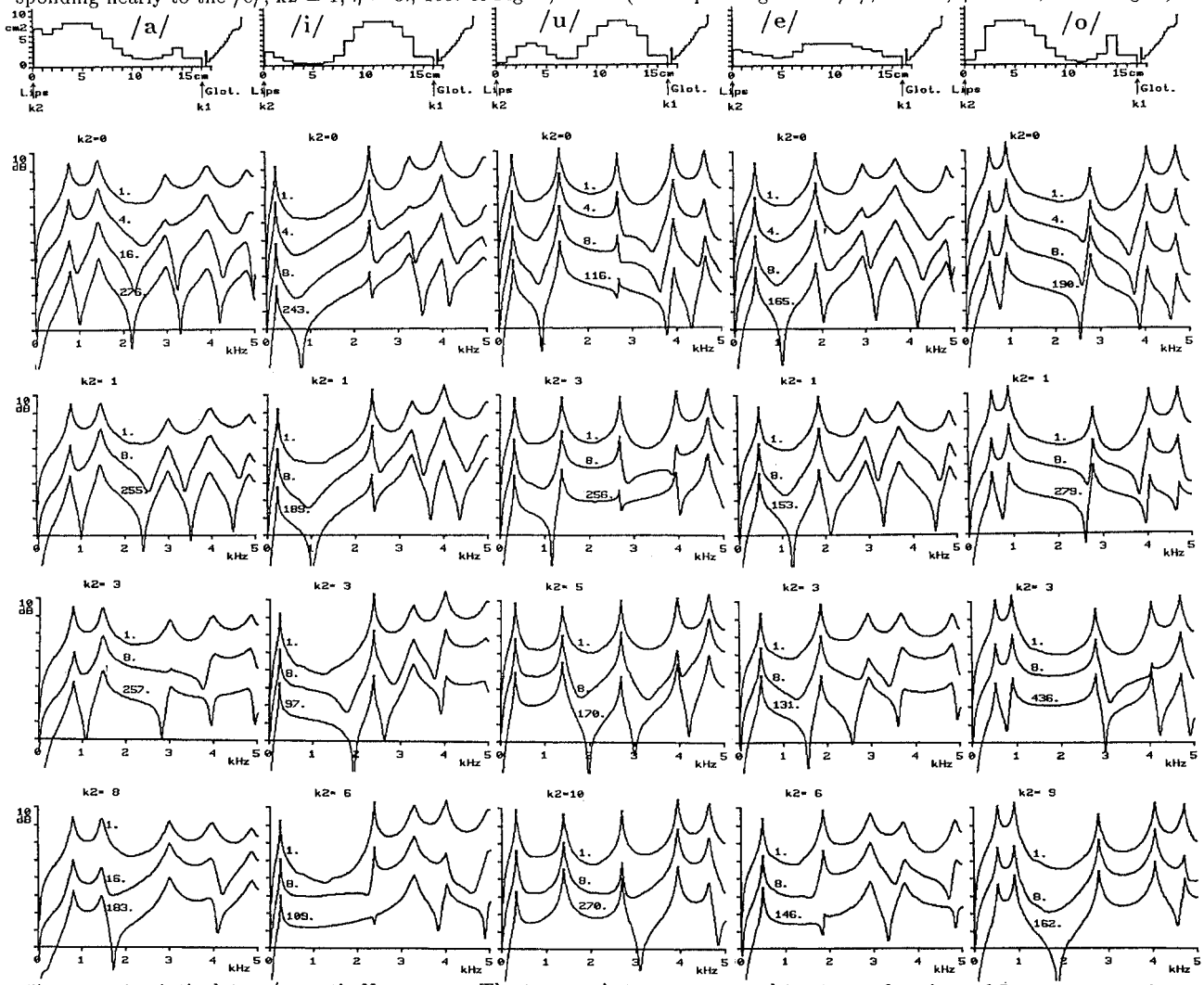

Fig. 6 An Articulatory/acoustic Nomogram. The top row is two-source vocal tract area functions of Japanese 5 vowels. To each vowel, the freq. transfer functions are shown in the column. where, k2 indicates 2nd source position from lips; 1st source position k1 is at the glottis; the parameter number above each transfer function is power ratio η of the 2nd source to the 1st .

The /o/ in Fig. 1 $\Rightarrow$ the /u/, k2=5, η =8.,170.
The /u/ in Fig. 2 $\Rightarrow$ the /e/, k2=1, η =8.,153.
The /a/ in Fig. 3 $\Rightarrow$ the /e/, k2=6, η =146.,8.
The /o/ in Fig. 4 $\Rightarrow$ the /e/, k2=3, η =8.,153.
The /a/ in Fig. 5 $\Rightarrow$ the /a/, k2=0, η =276.

Generalizing the above correspondences, the following knowledge are obtained:

1) The PSE of a vowel following a consonant closed by lips, tongue tip, tongue body (V in CV) corresponds to the two source vocal tract model in which the second source is placed at around the area of the lips, tongue tip, or tongue body, respectively. The intensity of zero gets weaker, as time passes from the end of the preceding close consonant to the following vowel center. This means in the model, that the second source intensity becomes weak compared to the glottal source in the transient part.

2) The PSE of a vowel preceding a closed consonant (V in VC) corresponds to the two source vocal tract model in which the second source is placed at about the position where the sound source is generated in the following closed consonant. The intensity of zero gets stronger as time lapses from the center of the vowel to the following closed consonant. It is, furthermore, pointed out that the vowel part zero positions reflect the articulatory position, such as, lips, tongue tip or tongue body used in the utterance of the nearest neighbouring closed consonants, respectively.

3) The PSE of a short vowel PSE, which is sandwiched between closed consonants, has a tendency to have such zeroes which are generated from the vowel vocal tract with sources placed at nearly the same source positions as generated in the preceding and following consonants.

A great percentage of vowels in continuous speech have neighboring closed consonants. As to the vowel in such phonemic context, the following hypothesis is proposed: "PSE variations such as zero growing, specific pole diminishing, or extreme bending of the over-all logarithmic power spectrum in the logarithmic frequency scale are created in the vowel. The reason is recognized in such a way that the vowel part is produced by a normal vocal tract with two sources: the first one is at the glottis and the second is configured as the direct effect of the neighbouring closed consonant. As a result, the second source position is nearly the same as that of the consonant. The intensity ratio of the second source to the first source increases as time passes from the vowel center to the following consonant, and decreases in the interval between the preceding consonant to the following vowel center".

The hypothesis leads to the conclusion that the continuous speech vowel part includes the nearest neighbouring closed consonant source position information: lips, tongue tip and tongue body. These are recognized to be dominant reasons why the continuous speech vowel PSE is enormously different from that of separately uttered sustained vowels.

4 Conclusion

We propose the hypothesis that the PSE of Japanese continuous speech vowels are created by a vocal tract with two-sources: the first at the glottis, the second at around the place corresponding to the neighbouring closed consonant's articulatory position.

This hypothesis leads to the conclusion that the continuous speech vowel part carries the information about the source position of the nearest neighbouring closed consonant: lips, tongue tip, or tongue body. These are recognized to be the dominant reason that why the continuous speech vowel PSE is enormously different from that of a separately uttered sustained vowel by the same speaker. Therefore, the model is expected to be valid for the systematic description of how the preceding and the following closed consonant information is compiled and superimposed onto the vowel part PSE.

The mechanism of the second source generation should be studied from the standpoint of science. As to the mechanism, it is highly probable that the consonant is articulatory constriction still remains or invades the vowel part when the consonant is a bi-labial or a tongue-tip one, or that the vowel articulatory constriction is supposedly much stronger in such a context compared to that in sustained or isolated spoken vowel, so that the second source is generated when the air flow pulse passes through the narrow channel after it passes through the glottis. In the second sound source generation process, the rapid movement of the articulatory organ configuring the constriction may contribute to preparing the easy sound generation condition.

It has been pointed out that the continuous speech vowel PSE may carry the neighbouring closed consonant articulatory position information, simultaneously. The view is coincident to the result of a human hearing test: that a short term wave-form of so-called vowel part cut out from continuous speech carries enough phonemic information of the neighbouring consonant in auditory sense. The proposed two-source vocal tract model can express the vowel duality that a vowel carries the consonantal aspects, simultaneously. Therefore, the model is expected to be valid for the basic model for future speech synthesis, and feature extraction for phoneme description for future continuous speech recognition.

References

[1] T.Nakajima and T.Suzuki,"Pitch pair synchronous PSE analysis method based on a non-steady state wave spectral model," Joun. of Acous. Soc. of Japan, Vol.44,No.11,900-908(1988)

[2] H.Ohmura and T.Nakajima,"Computation of consonantal vocal tract transfer functions in terms of reflection coefficients," Joun. of Acous. Soc. of Japan, Vol.46,No.12,18-27(1990)

[3] N.Umeda and R.Teranishi,"Phonemic quality and voice quality of Voice — Speech synthesis by acoustic vocal tract model— ," Joun. of Acous. Soc. of Japan, Vol.22,No.2,195-203(1966)