

Title: Discrimination of "Hot Potato Voice" Caused by Upper Airway Obstruction Utilizing a Support

Vector Machine

Authors:

Shintaro Fujimura M.D. ¹

Tsuyoshi Kojima M.D., Ph.D. ¹

Yusuke Okanou M.D. ¹

Kazuhiko Shoji M.D., Ph.D. ¹

Masato Inoue M.D., Ph.D. ²

Ryusuke Hori M.D., Ph.D. ¹

1. Department of Otolaryngology, Tenri Hospital, Tenri, Nara, Japan

2. Department of Electrical Engineering and Bioscience, School of Advanced Science

and Engineering, Waseda University, Shinjuku, Tokyo, Japan

Running title: Computational recognition of hot potato voice

Corresponding author:

Tsuyoshi Kojima, M.D., Ph.D.

200 Mishima-cho, Tenri, Nara 632-8552, Japan

Tel: +81-743-63-5611

Fax: +81-743-63-1530

E-mail: t_kojima@ent.kuhp.kyoto-u.ac.jp

This is the peer reviewed version of the following article: Discrimination of "Hot Potato Voice" Caused by Upper Airway Obstruction Utilizing a Support Vector Machine, which has been published in final form at Laryngoscope, 129:1301-1307, 2019, DOI:10.1002/lary.27584. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

ABSTRACT

Objectives

“Hot potato voice” (HPV) is a thick, muffled voice caused by pharyngeal or laryngeal diseases characterized by severe upper airway obstruction, including acute epiglottitis and peritonsillitis. To develop a method for determining upper-airway emergency based on this important vocal feature, we investigated the acoustic characteristics of HPV using a physical, articulatory speech synthesis model. The results of the simulation were then applied to design a computerized recognition framework using a Mel-frequency cepstral coefficient domain support vector machine (SVM).

Methods

Changes in the voice spectral envelope caused by upper airway obstructions were analyzed using a hybrid time-frequency model of articulatory speech synthesis. We evaluated variations in the formant structure and thresholds of critical vocal tract area functions that triggered HPV. The SVMs were trained using a dataset of 2,200 synthetic voice samples generated by an articulatory synthesizer. Voice classification experiments on test datasets of real patient voices were then performed.

Results

On phonation of the Japanese vowel /e/, the frequency of the second formant fell and coalesced with that of the first formant as the area function of the oropharynx decreased. Changes in higher-order formants varied according to constriction location. The highest accuracy afforded by the SVM classifier trained with synthetic data was 88.3%.

Conclusions

HPV caused by upper airway obstruction has a highly characteristic spectral envelope. Based on this distinctive voice feature, our SVM classifier trained using synthetic data was able to diagnose upper-airway obstructions with a high degree of accuracy.

Key Words

Hot potato voice, upper airway obstruction, articulatory speech synthesis, support vector machine

Level of Evidence

2c

INTRODUCTION

Certain pharyngeal or laryngeal diseases such as peritonsillitis or acute epiglottitis cause pathognomonic voice changes. The voice is described as thick or muffled, and is termed a “hot potato voice” (HPV), as if the patient was struggling with a mouthful of hot food. Emergency physicians and otolaryngologists are alert for this peculiar voice change because it implies that the patient has severe pharyngeal or laryngeal problems and is in danger of airway obstruction or suffocation. We have also encountered this change in patients suffering from upper airway obstruction caused by a large tumor or severe edema. Although HPV is well-recognized as an important sign of an upper airway emergency, very few studies have focused on this unique symptom, and its acoustic characteristics remain unclear ¹.

The source-filter model is a simplified speech production mechanism. In this model, speech is viewed as the passage of a glottal excitation signal (a source) through a time-varying linear filter that models the resonant characteristics of the vocal tract. Hence, it should be possible to represent the acoustic characteristics of HPV as distinct resonance characteristics caused by vocal tract constriction. We initially identified the common resonance characteristics, or spectrum envelope, of the voice, rendering it possible to identify patients with peritonsillitis or acute epiglottitis via voice assessment

alone. We then used computerized HPV recognition to establish a method for the diagnosis of dangerous upper airway obstructions. This is regarded as a classification problem, in which the classifier assigns input vectors (patient voices) to finite discrete categories (presence or absence of vocal tract constriction point). The classifier can be trained with previously categorized (voice) samples and labeled based on the corresponding vocal tract shapes. We used a supervised machine-learning technique (support vector machine; SVM) to this end.

However, the voice spectrum envelope reflects pronounced inter-individual variation in the phrase and morphology of the vocal tract. Such common variation may conceal subtle changes caused by pathological airway obstruction; thus it was necessary to accumulate and analyze numerous HPVs to reveal their common features. Supervised learning by SVMs requires a large training dataset. In practice, opportunities to examine and treat patients with severe upper airway obstructions are limited; thus, it has been difficult to define the features of HPV.

To resolve this problem, we used a model of physical, articulatory speech synthesis (Figure 1). Changes in the voice spectrum envelope caused by upper airway obstructions were calculated using this model and compared to those of affected patients; this revealed specific spectrum envelope characteristics of the unusual voice. Synthesized voices made by highly constricted vocal tracts were

thick and muffled, thus typical HPVs, and allowed effective analyses of acoustic characteristics.

Finally, we used large amounts of synthetic data to train SVMs to classify clinical HPVs of real patients with upper airway obstructions.

MATERIALS AND METHODS

I. Analyses of the acoustic characteristics of the HPV

A. The physical, articulatory speech synthesis model

The foundations for speech synthesis based on articulatory modeling were built by Fant (1960), Holmes et al. (1964), Flanagan (1972), Klatt (1976), and Allen et al. (1987). In the classic approach, wave propagation in the vocal tract is approximated by a digital ladder (or waveguide) filter operating in a discrete time domain. The celebrated Kelly-Lochbaum (KL) model² approximates the vocal tube as a set of cylindrical elements of overall length $L_e = c/F_s$, where c is the velocity of sound and F_s the audio sampling rate. However, differences in vocal tract shape or length cannot be easily accommodated using this model. We employed the hybrid time-frequency system^{3 4 5} of Sondhi and Schroeter, in which, a chain-matrix approach is used to compute the transfer function of the vocal tract in the frequency domain (Figure 1). This allows incorporation of realistic frequency-dependent losses

as well as variations in vocal tract length and the number of tubes that describe the vocal tract, independent of sampling frequency.

A chain matrix relates the pressure and volume velocity at the output side of a tube $[P_{out} \ U_{out}]$ to those at the input side $[P_{in} \ U_{in}]$ (capitalized to denote variables in the frequency domain obtained by performing Laplace transforms):

$$\begin{bmatrix} P_{in} \\ U_{in} \end{bmatrix} = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} \begin{bmatrix} P_{out} \\ U_{out} \end{bmatrix} = K \begin{bmatrix} P_{out} \\ U_{out} \end{bmatrix}$$

The elements of this matrix are easy to compute for a tube of uniform area, being specified to include wall vibration, viscous friction, etc. ⁴ In this study, an arbitrary tube was represented as a concatenation of elementary uniform sections. Considering non-nasal sounds for simplicity, the overall vocal tract chain matrix is the product of a sequence of elementary 2×2 matrices:

$$K = K_1 K_2 \dots K_N$$

Assuming that the tract is terminated by the radiation impedance Z_R of the lips, the transfer function of the entire vocal tract is:

$$\frac{U_{out}}{U_{in}} = \frac{1}{k_{21}Z_R + k_{22}}$$

This transfer function relates the volume velocity at the output and input sides of the vocal tract. The input impedance of the vocal tract, as seen at the glottis, is:

$$\frac{P_{in}}{U_{in}} = \frac{k_{11}Z_R + k_{12}}{k_{21}Z_R + k_{22}}$$

The frequency response of this model is obtained by substituting $s = i\omega$ for the transfer function. To calculate the transfer function of the entire vocal tract, we used the vocal tract variables of Sondhi et al.³ and the area function data of Story et al.⁶; these reflect the basic vocal tract shapes of five Japanese vowels (/a/, /i/, /u/ [a substitute for /ɯ/], /e/, and /o/), with vocal tract length $V_l = 17.0$ cm.

Next, we created a finite impulse response filter via inverse discrete Fourier transformation of the frequency transfer function. Applying this filter to a glottal source, prepared by linear-predictive-coding (LPC) inverse filtering of the lead author's voice, a final output voice sound was generated. All sound signal processing was performed at a sampling frequency of 44.1 kHz.

B. Acoustic characteristics of the synthesized HPV and its auditory impressions

The primary aim of this study was to investigate the relationship between the change in voice spectral envelope caused by vocal tract constriction and any association with HPV. First, we used the above-described physical simulation model to analyze gradual variations of the spectral envelope caused by vocal tract constriction. Next, we generated five test datasets consisting of around 400 synthetic vowel sounds each; we gradually decreased the vocal tract area function of 11 elementary uniform sections of the oropharynx (from the top of the epiglottis to the velum) beginning with the normal values and

decreasing to a minimum of 0.01 cm². We varied the area function decreasing intervals slightly and randomly by every voice sample production, to uniformly distribute the sampling points of the vocal tract shape. The number of voice samples therefore differed in each dataset. Five otolaryngologists evaluated each sound and noted whether it sounded such like an HPV. The labeled voice samples were scatter-plotted in terms of their formant-frequency space (the first $[F_1]$ to fourth $[F_4]$ formants), and then the threshold of the critical area function triggering the HPV was obtained.

II. Computational recognition of the HPV

Based on the acoustic characteristics of the HPV revealed by the articulatory speech synthesizer, we designed a computerized HPV recognition system using a Mel-Frequency Cepstral Coefficients^{7 8} (MFCC)-domain SVM.

A. Feature extraction of HPV

When building a computerized HPV recognition framework, voice spectra must be transformed to feature vectors suitable for use in a pattern-recognition algorithm. MFCCs are most commonly used to extract features employed in automatic speech recognition, and were first used to this end by Bridle and Brown in 1974^{7 8}. MFCCs mimic the logarithmic perception of the loudness and pitch of the human auditory system, rendering it possible to efficiently reduce the dimensionality of spectrum

envelope information. We extracted each voice sample over time units of 100 ms and converted them into MFCC feature vectors using 20 triangular filters and 12 cepstral coefficients, excluding the first coefficient.

B. Classification of clinical HPV using a SVM

SVM is a powerful discriminative classifier pioneered by Vapnik ⁹. It SVM maps an input onto a high-dimensional space and then defines a maximum-margin hyperplane that optimally separates the data with high generalizing capability and performance. We employed a C-SVM algorithm with a radial basis function kernel.

Using the above physical simulation model, we created the training dataset for SVMs, containing 2,200 synthetic voice samples with different vocal tract area functions for 11 elementary uniform sections of the oropharynx (from the top of the epiglottis to the velum), from the normal values to a minimum of 0.01 cm². The sustained Japanese vowel /e/ was chosen for the analyses because its normal phonation requires a relatively large oropharyngeal space. It was therefore assumed that constrictive changes in the oropharynx would more extensively affect phonation of this vowel than other vowels. Prior to the SVM training phase, synthetic voice samples of the training dataset were classified and labeled in terms of vocal tract constriction at an arbitrary minimum area function

threshold value in the sections of the whole oropharynx (thus not by any auditory impression of an HPV). To explore the effects of the various extents of vocal tract constriction, the SVMs were trained using several training datasets with different area function thresholds (from 0.15 cm² to 0.30 cm²). Consequently, the SVMs were able to discriminate the voices of patients with vocal tract constriction in some part of the oropharynx based on arbitrarily set thresholds.

The test dataset included 12 HPV samples from real patients with upper airway obstructions (Table 1) and six voice samples from patients with no constrictive upper airway disease. To reduce analytical inaccuracies caused by the low amount of available test data, we created five test datasets in which the offset points of the cropping analyses were randomly changed; the average of the values from the five datasets served as the final result.

The SVM hyperparameters were tuned via eight-fold cross-validation with the training datasets and then fine-tuned using the test dataset. Next, we performed voice classification experiments on the test datasets employing the trained SVMs. All training, tuning, and classification experiments employed the R-package e1071 (Mayer, 2014), which wrapped libsvm (Chang and Lin, 2011).

RESULTS

I. Acoustic characteristics of the HPV

Figure 2 depicts the trajectory of the simulated spectrum envelope when articulating the Japanese vowel /e/, caused by gradual decreases in the area functions of elements corresponding to three different parts of the pharynx (the top of the arytenoid, the top of the epiglottis, and around the palatine tonsil). The volumes range from the normal values (blue) to 0.01 cm² (red). Below, we focus on formant variation representative of changes in the spectral envelope. In general, the F_2 frequency gradually fell as the area function decreased, to initially approach and finally coalesce with F_1 at about 1,000 Hz. However, changes in the higher-order formants varied by the location of the constriction, which at the top of the arytenoid was associated with relatively good preservation of higher-order formant structure (Figure 2a); the resonance characteristic resembled that of a single cylindrical pipe. Constriction at the top of the epiglottis decreased the F_5 frequency, which ultimately became integrated with F_4 (Figure 2b). Constriction around the palatine tonsil decreased the F_6 frequency, which ultimately became integrated with F_5 , and also reduced the F_4 frequency (Figure 2c). Such variation in higher-order formants subtly affected the timbre of the HPV.

We found that the variation in formant structures associated with the different vocal tract constriction patterns were very complicated; we do not present the details for all vowels or constriction

patterns. We show typical changes in formant structure associated with phonation of other Japanese vowels in Figure 3.

Figure 4 shows the threshold values of the area functions at which simulated voices were perceived as HPVs for 11 elementary uniform sections corresponding to the oropharynx (ranging from the top of the epiglottis to the velum). The average threshold (all sections) was 0.15 cm^2 .

II. Computational recognition of the HPV

Figure 5 shows the testing dataset classification results and the efficiencies of interpretation by trained SVMs; the different threshold values of area function used to assess the synthesized voice samples of the training datasets are depicted. The highest efficiency was 88.3% at a threshold of 0.23 cm^2 . As the threshold increased, the false-positive rate rose, and as the threshold decreased, the false-negative rate rose, as expected given the classification framework employed.

DISCUSSION

HPV is a well-known symptom of a peritonsillar abscess or acute epiglottitis and is a recognized warning sign of upper airway obstruction. However, both the details of the voice and the mechanism of its generation remain unclear. Finkelstein et al. found that peritonsillar infection immobilized the palatopharyngeal, superior constrictor, and levator velopalatini muscles of the affected side, and that

HPV reflected an underlying, transient velopharyngeal insufficiency combined with muffled oral resonance¹⁰. However, we sometimes encounter patients with HPV or similarly altered voices who have upper airway obstructions caused by conditions other than a peritonsillar abscess (acute epiglottitis, large epiglottic cyst¹¹, laryngopharyngeal edema, or a large pharyngeal tumor¹²). Therefore, we consider that changes in vocal tract shape or wall surface properties per se, rather than velopharyngeal insufficiency, alter the resonance characteristics of the vocal tract to generate the HPV; our simulation results showed that the sound generated by a simple vocal tract, part of which was heavily constricted, was actually perceived as a thick and muffled HPV. The average threshold required to change a plain voice to a HPV was 0.15 cm² in our simulation, representing an extremely severe airway constriction, substantiating the pathological significance of such a voice (which constitutes a clinical emergency). In terms of velopharyngeal insufficiency, in another simulation of narrow-conjunction nasal tract coupling (achieved by tapping two special matrices at the velum in a chain matrix⁴), imitating velopharyngeal insufficiency, we found that it introduces zero above F_1 and formed a distinctive envelope shape around 1,000 Hz, resembling the changes caused by oropharyngeal constriction (Figure 6). This shape change appeared to accentuate the HPV caused by oropharyngeal constriction.

Our articulatory synthesis model has several limitations. For simplicity, we did not seek to emulate the surface properties of the vocal tract wall and we ignored the transient dynamics of linguistic activity. We assumed that these did not contribute to the acoustic characteristics of the HPV because our simplified model well-represented the perceptual auditory features of that voice; refinements can be made if the simulation objectives of other implementations differ.

Our approach allows evaluation of the resonance characteristics of HPV using formant frequencies, which is an easily comprehensible method to evaluate the voice spectral envelope. As is evident from the above discussion, when simply combined with a linear classifier, the formant frequencies can be used to construct a machine learning model enabling HPV discrimination. However, several problems with this design must be resolved before the method can be put into practice. First, the characteristics of the voice spectral envelope are represented not only by the formant frequencies, but also by the bandwidth or energy of each formant. Secondly, as previously explained, some of the formant frequencies were close to each other and finally coalesced into a single spectral peak as constriction of the vocal tract became severe. Therefore, even if a sequence of formant frequencies is calculated using real voice data, it will be difficult to ensure one-to-one correspondences between the components of this formant sequence and those of other real or synthetic data. For these reasons, the

characteristic features used for a machine learning classifier should be representative of the entire structure of the voice spectral envelope. MFCC is a reasonable feature that is able to describe the overall shape of the spectral envelope in a small set of coefficients.

In addition, and importantly, we utilized synthetic data generated by an articulatory synthesizer to train the HPV classifier and supervise learning. Obviously, even the best machine-learning model is useless if high-quality data are lacking. Too few clinical HPV samples are available to allow supervised machine-learning. However, the articulatory speech synthesizer generated a very large number of voice samples labeled with arbitrary thresholds. The use of only virtually synthetic data to train machine-learning models may initially appear to be inappropriate, although synthesized and real HPVs sound very similar. On careful consideration, however, supervised learning using synthetic data affords several advantages. Synthetic data can be labeled with perfect accuracy as long as the model precisely represents the required phenomena, even when such data are very difficult to capture or assess in real life.

The excellent classification results that we obtained may imply that our simulation of articulatory synthesis well mimics the real HPV phonation phenomenon. In the current study, the different constriction positions in the oropharynx could not be discriminated because too few test data were

available. Our framework of machine-learning models trained by synthetic data can also be applied to several other clinical issues in pathological voice diagnosis, such as a nasal voice or dysarthria. Constructing a high-quality dataset with abundant real voice data will enable an improved diagnostic capability.

CONCLUSIONS

Using a physical, articulatory speech synthesis model, we showed that HPV caused by upper airway obstruction has a highly characteristic spectral envelope. On the basis of the distinctive voice features of HPV, the SVM classifier, trained using synthetic data, afforded excellent efficiency in terms of the diagnosis of upper airway obstruction during examination of the patient's voice.

References

1. Bhutta MF, Harries ML, Worley GA. "Hot Potato Voice" in Peritonsillitis: A Minomer. *Journal of Voice*. 2006;20(4):616-622.
2. Kelly LJ, Lochbaum CC. Speech synthesis. Paper presented at: Proceedings of the Fourth International Congress on Acoustics, 1962; Copenhagen.
3. Sondhi MM. Model fo wave propagation in a lossy vocal tract. *The Journal of the Acoustical Society of America*. May 1974;55(5):1070-1075.
4. Sondhi MM, Schroeter J. A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. July 1987;35(7):955-967.
5. Sondhi MM, Schroeter J. Speech Production Models and Their Digital Implementations. *Digital Signal Processing Handbook*: CRC Press LLC; 1999.
6. Story BH, Titze IR. Vocal tract area functions from magnetic resonance imaging. *The Journal of the Acoustical Society of America*. July 1996;100(1).
7. Bridle JS, Brown MD. An Experimental Automatic Word-Recognition System. Ruislip, England: Joint Speech Research Unit; 1974. 1003.

8. Davis SB, Mermelstein P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing. August 1980;28(4).
9. Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995;20(3):273-279.
10. Finkelstein Y, Bar-Ziv J, Nachmani A, Berger G, Ophier D. Peritonsillar Abscess as a Cause of Transient Velopharyngeal Insufficiency. Cleft Plate Craniofacial Journal. July 1993;30(4):421-428.
11. Fang TJ, Cheng KS, Li HY. A Huge Epiglottic Cyst Causing Airway Obstruction in an Adult. Chang Gung medical journal. April 2002;25(4):275-8.
12. Nicoli TK, Gupta S, Kotecha B. Radiofrequency thermotherapy in the treatment of tongue base lymphoid hyperplasia: case report. The Journal of laryngology and otology. March 2015;129(4):398-402.

Table 1. Causes of upper airway obstruction associated with “hot potato voice” (HPV) in test patients.

Causes of upper airway obstruction	No.
Acute epiglottitis	3
Oropharyngeal cancer	3
Malignant lymphoma of the root of the tongue	3
Laryngopharyngeal edema with Ludwig's Angina	1
Large thyroid tumor	1
Cicatricial stenosis of the oropharynx	1

Figure 1. The workflow with a hybrid, time-frequency domain model of articulatory speech synthesis, and the normal vocal tract shape associated with phonation of the Japanese vowel /e/.

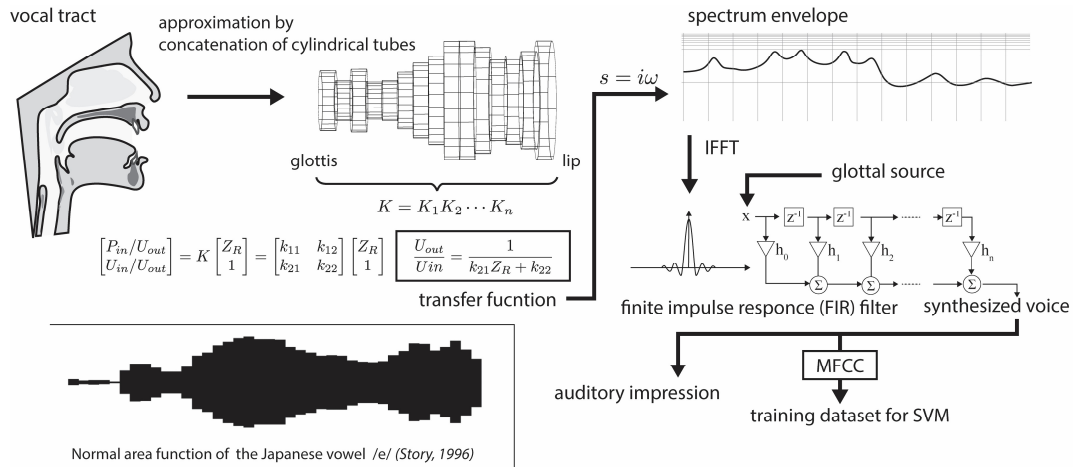


Figure 2. Spectrum envelope changes in the Japanese vowel /e/ caused by vocal tract constriction.

a) Constriction at the top of the arytenoid decreases the frequency of F_2 and integrates F_2 with F_4 .

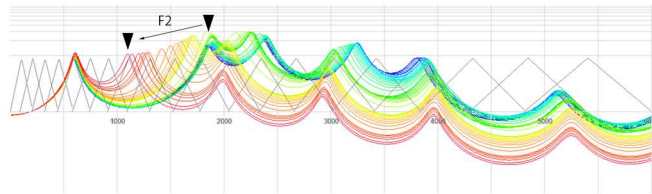
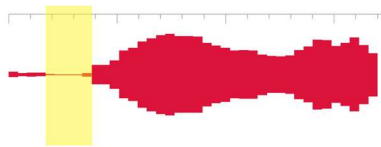
b) Constriction at the top of the epiglottis decreases the frequency of F_5 and integrates F_5 with F_4 ,

in addition to the changes described in a) above.

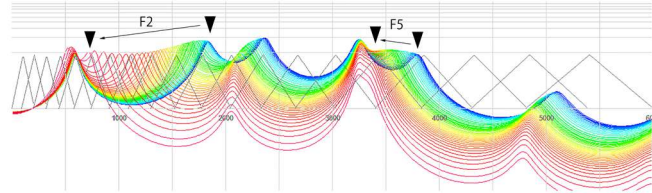
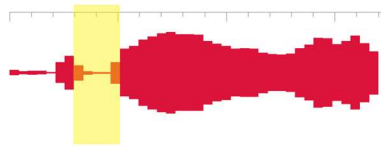
c) Constriction around the palatine tonsil decreases the frequency of F_6 and integrates F_6 with F_5

in addition to the changes described in a) above.

a) Constriction at the top of the arytenoid



b) Constriction at the top of the epiglottis



c) Constriction around the palatine tonsil

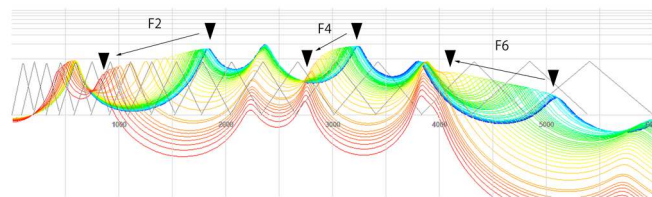
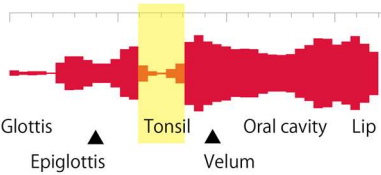


Figure 3. Formant structure changes caused by upper airway obstruction.

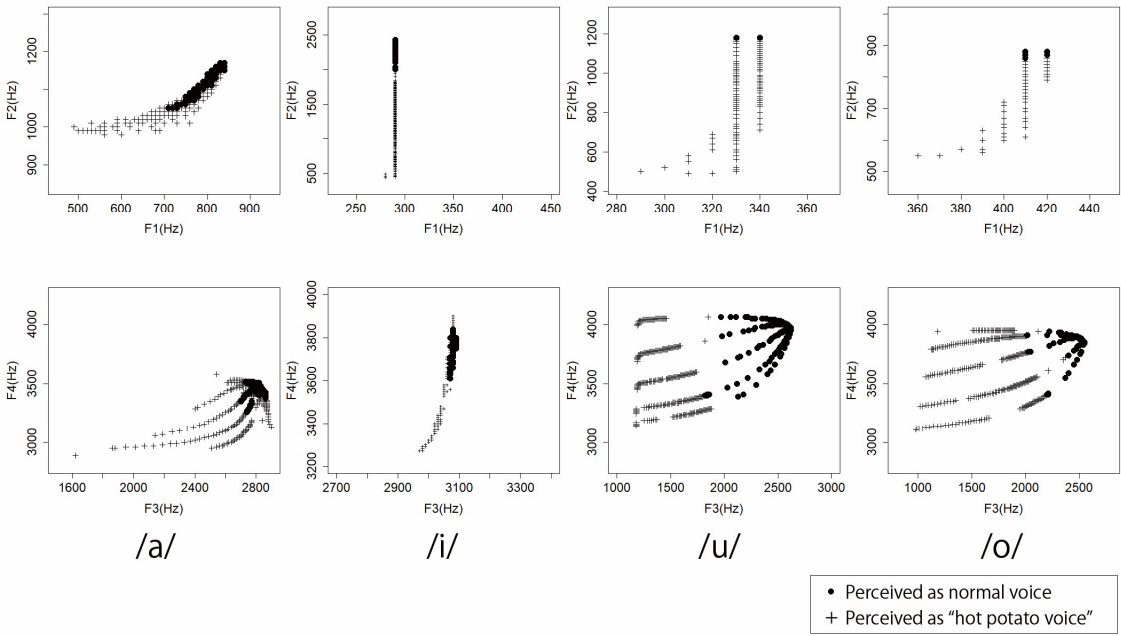


Figure 4. Figure 4. The threshold values of the area functions at which simulated voices were perceived as HPVs at 11 elementary uniform sections corresponding to regions of the oropharynx.

a) The threshold values depicted with the normal values of the area functions of the oropharynx.

b) The details of the threshold values of the area functions triggering the HPV.

The threshold values of the area function at which the simulated voices were perceived as “hot potato voice” for eleven elementary uniform sections corresponding to the oropharynx.

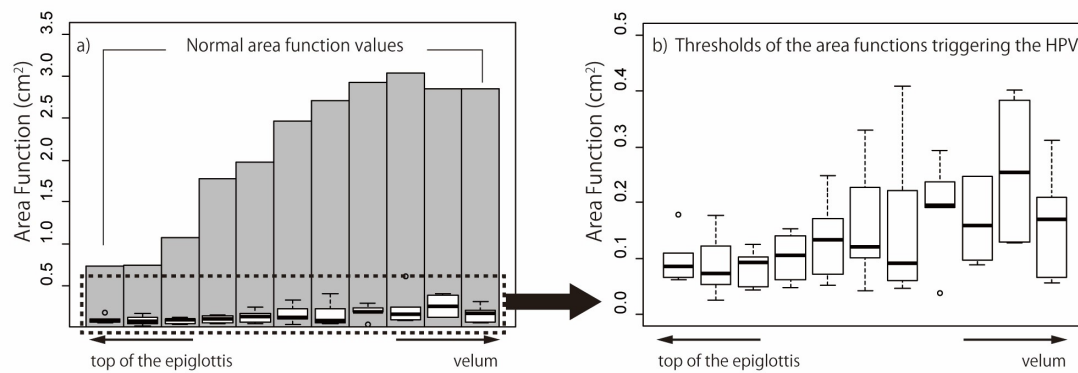


Figure 5. The results of classification of patient voices and the efficiency of such classifications performed using trained support vector machines (SVMs). Various thresholds are shown. The bar lengths represent (from left to right) the proportions of correctly predicted negative samples, false-positive samples, false-negative samples, and correctly predicted positive samples.

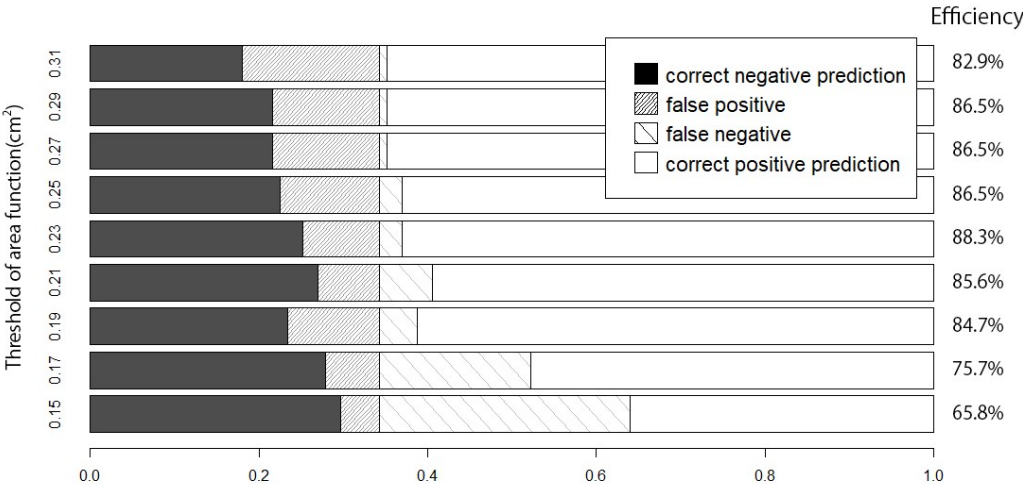


Figure 6. Simulated spectral envelope changes caused by narrow-conjunction nasal coupling. The blue line shows the spectral envelope of the Japanese vowel /e/ and the red line the spectral envelope in the presence of nasal tract coupling. The arrowhead indicates the zero point of nasal tract coupling.

