

Comparison of Multivariate Analysis Methods as Applied to English Speech

Zhang, Yixin
Human Science International Course, Kyushu University

Nakajima, Yoshitaka
Sound Corporation

Ueda, Kazuo
Department of Human Science, Faculty of Design, Kyushu University

Kishida, Takuya
Graduate School of Informatics and Engineering, The University of Electro-Communications

他



<https://hdl.handle.net/2324/7153265>

出版情報 : Applied Sciences. 10 (20), pp.7076-, 2020-10-12. Multidisciplinary Digital
Publishing Institute : MDPI
バージョン :
権利関係 : Creative Commons Attribution 4.0 International



Article

Comparison of Multivariate Analysis Methods as Applied to English Speech

Yixin Zhang ^{1,*}, Yoshitaka Nakajima ², Kazuo Ueda ³ , Takuya Kishida ⁴ and Gerard B. Remijn ⁵ 

¹ Human Science International Course, Kyushu University, Fukuoka 815-8540, Japan

² Sound Corporation, Fukuoka 813-0001, Japan; yoshitaka.nakajima@100years.life

³ Department of Human Science, Faculty of Design/Research Center for Applied Perceptual Science/Research and Development Center for Five-Sense Devices, Kyushu University, Fukuoka 815-8540, Japan; ueda@design.kyushu-u.ac.jp

⁴ Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan; kishida@uec.ac.jp

⁵ Department of Human Science, Kyushu University/Research Center for Applied Perceptual Science, Kyushu University, Fukuoka 815-8540, Japan; remijn@design.kyushu-u.ac.jp

* Correspondence: 3DS20005M@s.kyushu-u.ac.jp

Received: 20 August 2020; Accepted: 5 October 2020; Published: 12 October 2020



Abstract: A newly developed factor analysis, origin-shifted factor analysis, was compared with a normal factor analysis to analyze the spectral changes of English speech. Our first aim was to investigate whether these analyses would cause differences in the factor loadings and the extracted spectral-factor scores. The methods mainly differed in whether to use cepstral liftering and an origin shift. The results showed that three spectral factors were obtained in four main frequency bands, but neither the cepstral liftering nor the origin shift distorted the essential characteristics of the factors. This confirms that the origin-shifted factor analysis is more recommendable for future speech analyses, since it would reduce the generation of noise in resynthesized speech. Our second aim was to further identify acoustic correlates of English phonemes. Our data show for the first time that the distribution of obstruents in English speech constitutes an L-shape related to two spectral factors on the three-dimensional configuration. One factor had center loadings around 4100 Hz, while the other was bimodal with peaks around 300 Hz and 2300 Hz. This new finding validates the use of multivariate analyses to connect English phonology and speech acoustics.

Keywords: speech perception; English phoneme; factor analysis; spectral change; sonority

1. Introduction

The frequency range covered by the human hearing system is remarkable. In an initial attempt to model the hearing system's frequency resolution, Zwicker (1961) proposed the bark scale, which divides the audible frequency range into 24 frequency ranges between 20 to 15500 Hz [1]. These 24 frequency ranges are now commonly considered as representing 24 "critical bands", which can be regarded as a series of bandpass filters through which sounds are processed [2]. The frequency range of speech as in a common AM-broadcasting system is limited to a relatively narrow range of approximately <7000 Hz, and in classic models it is often represented by 20 critical bands between 50 and 7000 Hz in the auditory periphery [3]. In order to resynthesize fairly intelligible speech, however, even fewer frequency bands seem sufficient.

Previous studies related to this proceeded from a series of principal component analyses and listening experiments. Plomp et al. (1967) found that only two main principal components could

express 70% of the phonological features of 15 Dutch vowels [4]. Moreover, the first and the second principal component of these vowels were found to be distinctly related to the formants of the vowels [5]. In a later study with principal component analysis, it was found that 3–5 components might convey enough clues to represent continuous speech [6].

In one of the most recent comparative studies using factor analysis with eight different languages/dialects, Ueda and Nakajima (2017) showed that three spectral factors of 20-critical-band-filtered speech universally appeared in all speech samples they used [7]. Interestingly, Nakajima et al. (2017) further found that one of the three common spectral factors extracted from English speech, which appeared in a frequency range around 1100 Hz, was highly related to sonority or aperture, terms used in phonology [8]. Sonority has been a hot topic for more than a century in speech research. It has played an important role in understanding syllable formation; many linguists believe that the structure of a syllable is driven by the principle of sonority [9]. Sonority is a concept used to classify phonemes, in the following order from the highest to the lowest on the sonority scale: vowels, glides, liquids, nasals, fricatives/affricates, and plosives [10]. The highest sonority is attributed to the phoneme that constitutes the closest clue to the core in English syllables [11]. Nakajima et al. (2017) divided all English phonemes observed in a speech database into three major categories from high to low sonority, i.e., vowels, sonorant consonants (glides, liquids, nasals), and obstruents (fricatives/affricates, plosives). Their factor analysis showed that a factor related to the frequency range around 1100 Hz was specifically associated with vowels and sonorant consonants. By contrast, another spectral factor with a frequency range above 3300 Hz was strongly related to obstruents. Based on this, it was suggested that the extraction of spectral factors can be a powerful tool to connect phonology and acoustic aspects of English speech [8].

It was problematic that the subspace selected in the factor analysis of Nakajima et al. (2017) did not include the acoustically silent point of the original spectral space. In this case, if speech were to be resynthesized, it would contain positive power at any temporal point (in most cases). This would make it unsuitable for listening experiments, because the listeners would perceive a steady background noise in the resynthesized speech, which then would not contain any silent part. To avoid this problem, Kishida et al. (2016) developed a new factor analysis method, based on what they called the “origin-shifted principal component analysis” [12]. In this method, the starting point for calculating eigenvectors was moved to the acoustically silent point, instead of the gravity center of the data points. This reduced the generation of the noise and improved resynthesized speech. Furthermore, they confirmed that there are indeed three or four common factors in spoken British English, Japanese, and Mandarin Chinese. If the number of spectral factors used for speech resynthesis is three or more, the basic language information will be retained in the noise-vocoded speech. However, it is still necessary to determine in detail how different the results of the normal factor analysis [8] and the modified factor analysis, i.e., origin-shifted factor analysis [12], can be when applied to English speech. In the present study, we used four different factor analyses for this purpose.

Another purpose of this study was to determine the acoustic correlates of English phonemes by using the aforementioned factor analyses. We specifically focus on how English obstruents appear in the obtained factor spaces. Using normal factor analysis, Nakajima et al. (2017) [8] suggested that obstruents are almost exclusively represented on a factor related to a frequency range above ~3300 Hz. However, it needs to be confirmed whether the same will be indicated using the origin-shifted factor analysis, following Kishida et al. (2016) [12]. The origin-shifted factor analysis, which always selects a subspace including the silent point, should be more sensitive to spectral changes near the silent point, and thus may give new insight into the acoustic natures of relatively weak sounds such as obstruents. Since an obstruent always forms the beginning or end of a syllable in English, sometimes with a consonant cluster, to determine acoustic correlates of obstruents is highly important. Already much is known about sonorous phonemes, like vowels, but less is known about the acoustics related to phonemes lower in the sonority hierarchy (cf. [10]).

2. Methods

2.1. Speech Samples

To compare the results of the origin-shifted factor analysis with the results of the normal factor analysis, the same 200 English sentences as in Nakajima et al. (2017) [8] were used. The sentences were spoken by one male and two female native speakers of English, taken from “The ATR British English Speech Database” [13]. The speech samples were recorded with a sampling frequency of 12000 Hz and 16-bit linear quantization. The speech signals of all the spoken sentences in the database were segmented into individual phonemes and were labeled utilizing the Machine Readable Phonetic Alphabet (MRPA) [14]. A total number of 31663 “phonemes” were labeled, but to maintain consistency with Nakajima et al.’s study (2017) [8], in which the phoneme labeling of this database had to be reexamined, 7523 were omitted, and the same 24140 English phonemes were taken up as analysis samples.

2.2. Procedure

Following the Nyquist theorem [15] and the 12,000-Hz sampling frequency of the English speech samples, 19 critical-band filters were constructed to cover the frequency range of 50–5300 Hz. Their center frequencies ranged from 75–4800 Hz. The frequency bands were adopted from Zwicker and Terhardt (1980) [16]. The frequency range below 50 Hz was not utilized because it should have been less relevant to the speech signal.

The factor analyses were performed as follows. Nakajima et al. (2017) calculated the powers of the filter outputs by squaring and moving-averaging them with a Gaussian window of $\sigma = 5$ ms. Thus, they obtained a smoothed power fluctuation for each filter, and the powers for the 19 filters were sampled at every 1 ms as 19 variates for factor analysis. In comparison with that method, here we followed Kishida et al. (2016) [12], where the speech signals were sampled every 1 ms with a 30-ms-long Hamming window. Kishida et al. then transferred every 30-ms-long segment to a power spectrum by Fast Fourier Transform (FFT) [12]. Following this, the power spectrum was smoothed with a 5-ms short-pass lifter by cepstral analysis [17]. The origin of the factors, which are represented as orthogonal vectors in the 19-dimensional variate space, was shifted from the gravity center of the data points to the silent point, at which all variates (powers) are zero. In this origin-shifted factor analysis, a silent part in the speech signal is represented always at the silent point. This method thus should be suitable for observing the natures of relatively weak sounds, which are closer to the silent point. This was not the case in Nakajima et al. (2017) [8], in which a normal factor analysis was used.

Together, we here performed four different factor analyses: with and without cepstral liftering, and with and without the origin shift. We obtained the factor loadings and took up the factor scores of the central midpoints of time for the labeled phonemes. We then analyzed the distribution of the phonemes as represented in the three-dimensional factor space. The three-dimensional axes derived from the four factor analyses were rotated by varimax rotation [18], resulting in spectral factors.

3. Results

3.1. Factor Loadings of the Three Spectral Factors

From the perspective of the analysis processes, the major differences between the two methods, i.e., the factor analyses in Nakajima et al. (2017) [8] and in Kishida et al. (2016) [12], were whether to use the cepstral liftering and the origin shift to the silent point. Figure 1 shows the factor loadings for all the English speech samples spoken by the three native speakers, obtained with each of the four analysis methods. One division of the horizontal axis corresponds to 0.5 critical bandwidth. Four main frequency bands were obtained. The first constituted a low-frequency band, approximately from 50 to 600 Hz. The second was a mid-low frequency band, approximately from 600 to 1700 Hz, followed by a mid-high-frequency band, approximately from 1700 to 3000 Hz. Finally, the fourth range was a

high-frequency band, approximately above 3000 Hz. Confirming the results of Ueda and Nakajima (2017) [7], Nakajima et al. (2017) [8], and Kishida et al. (2016) [12], these four frequency bands were related to three spectral factors. One factor, the “low & mid-high factor” (Figure 1, red line) was bimodal in that the frequencies of relatively high loadings were around 300 Hz and around 2300 Hz. The second factor, the “mid-low factor”, was located around 1100 Hz (Figure 1, black line). The third factor, the “high factor”, was located around 4100 Hz (Figure 1, blue line).

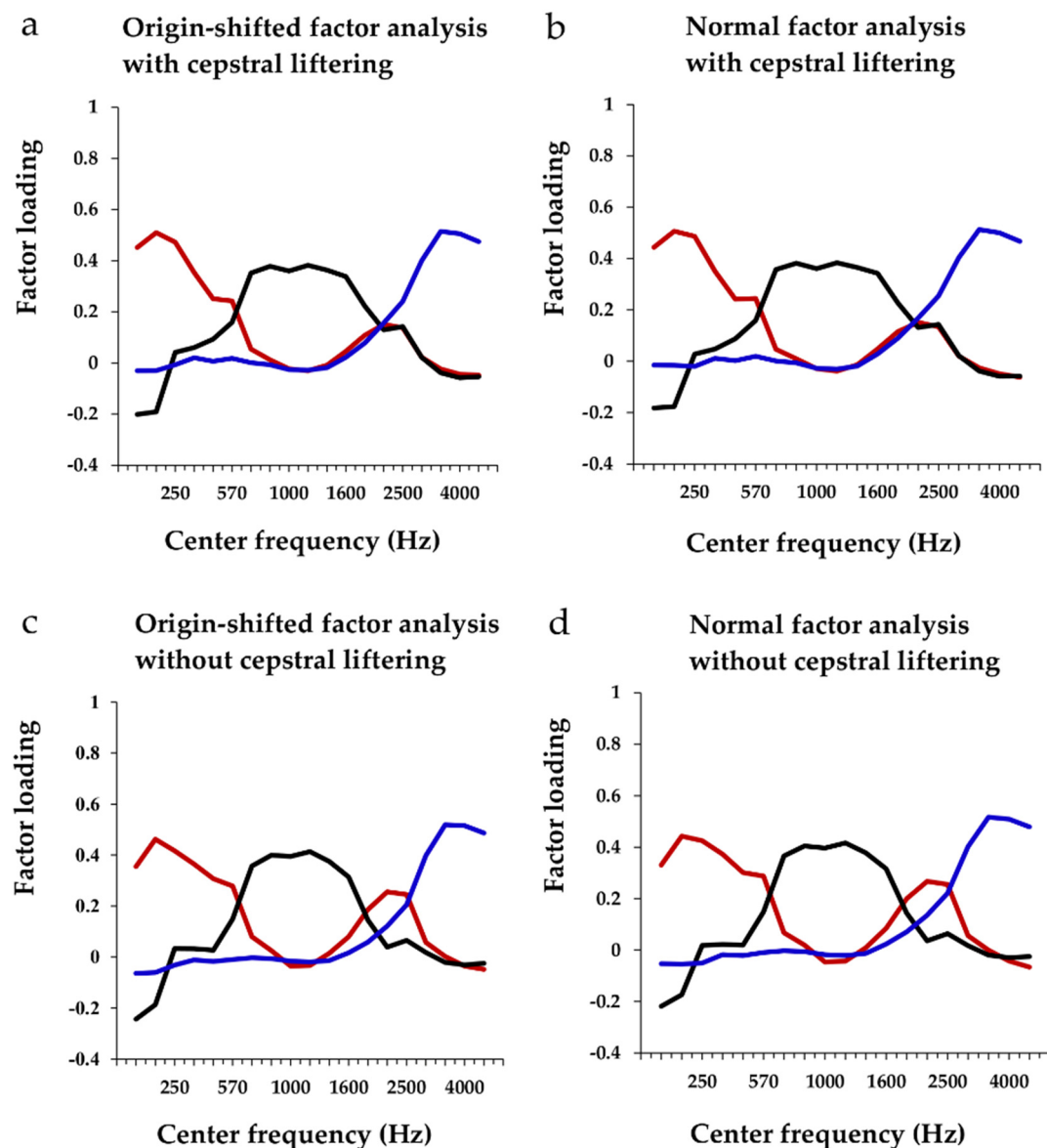


Figure 1. Factor loadings of the three extracted spectral factors of 200 English speech samples from three native speakers. Four different factor analyses were performed: (a) origin-shifted factor analysis with cepstral liftering, following [12]; (b) normal factor analysis with cepstral liftering; (c) origin-shifted factor analysis without cepstral liftering; (d) normal factor analysis without cepstral liftering.

The cumulative contributions of the three spectral factors were around 43% in Figure 1a,b, while they were around 45% in Figure 1c,d. As can be seen by comparing Figure 1a,b, or by comparing Figure 1c,d, the origin shift did not greatly affect the factor loadings—the factor loadings with origin shift and without origin shift turned out to be very similar. Cepstral analysis, however, affected the factor loadings for just one factor. That is, the second peak of the bimodal “low & mid-high factor”

was prominent without cepstral liftering, as can be seen in Figure 1c,d, but was reduced with cepstral liftering, as shown in Figure 1a,b. Thus, the similarity in the factor loadings between Figure 1a,b or between Figure 1c,d confirmed that the origin shift keeps the essential features of the factors in the analyses.

3.2. Factor Scores of the Three Spectral Factors

Following the analysis of the factor loadings, the factor scores were analyzed. In Figures 2–5 below, panels (a), (b), and (c) show the distributions of combinations of the three spectral-factor scores, i.e., between the “low & mid-high factor” (around 300 and 2300 Hz), the “mid-low factor” (around 1100 Hz), and the “high factor” (above 3000 Hz). Panel (d) in Figures 2–5 shows the distribution in the three-dimensional configuration as viewed in a direction from above-right to below-left in panel (a). In panel (d), the horizontal axis represents the combination of the “mid-low factor” and the “high factor”, by calculating $(x-y)/\sqrt{2}$, where x indicates the factor score of the “mid-low factor”, and y the factor score of the “high factor”, following Nakajima et al. (2017) [8].

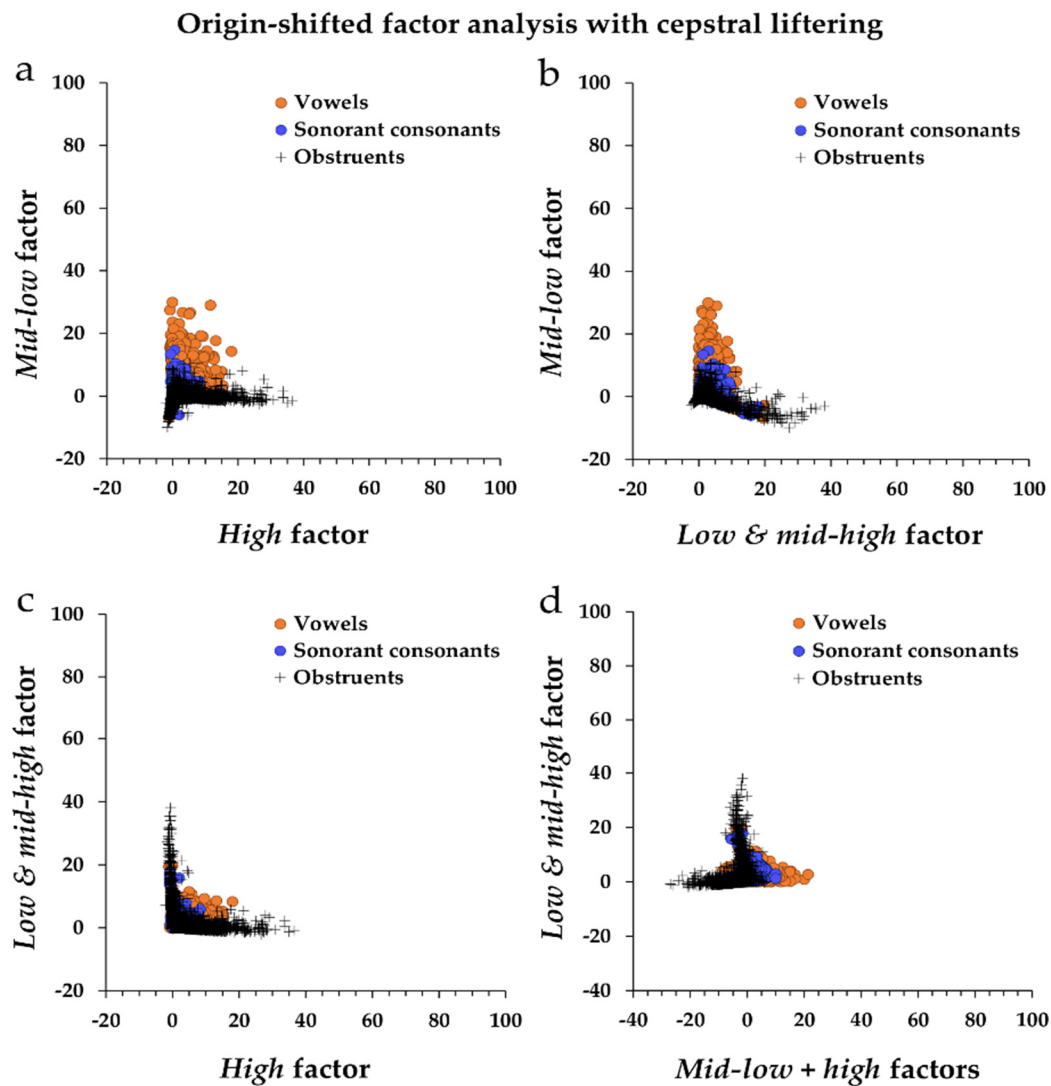


Figure 2. The distribution of the factor scores by origin-shifted factor analysis with cepstral liftering, following Kishida et al. (2016) [12]. (a–c) represent all possible combinations of two factors out of the three factors, and (d) shows a two-dimensional mapping of the three factors.

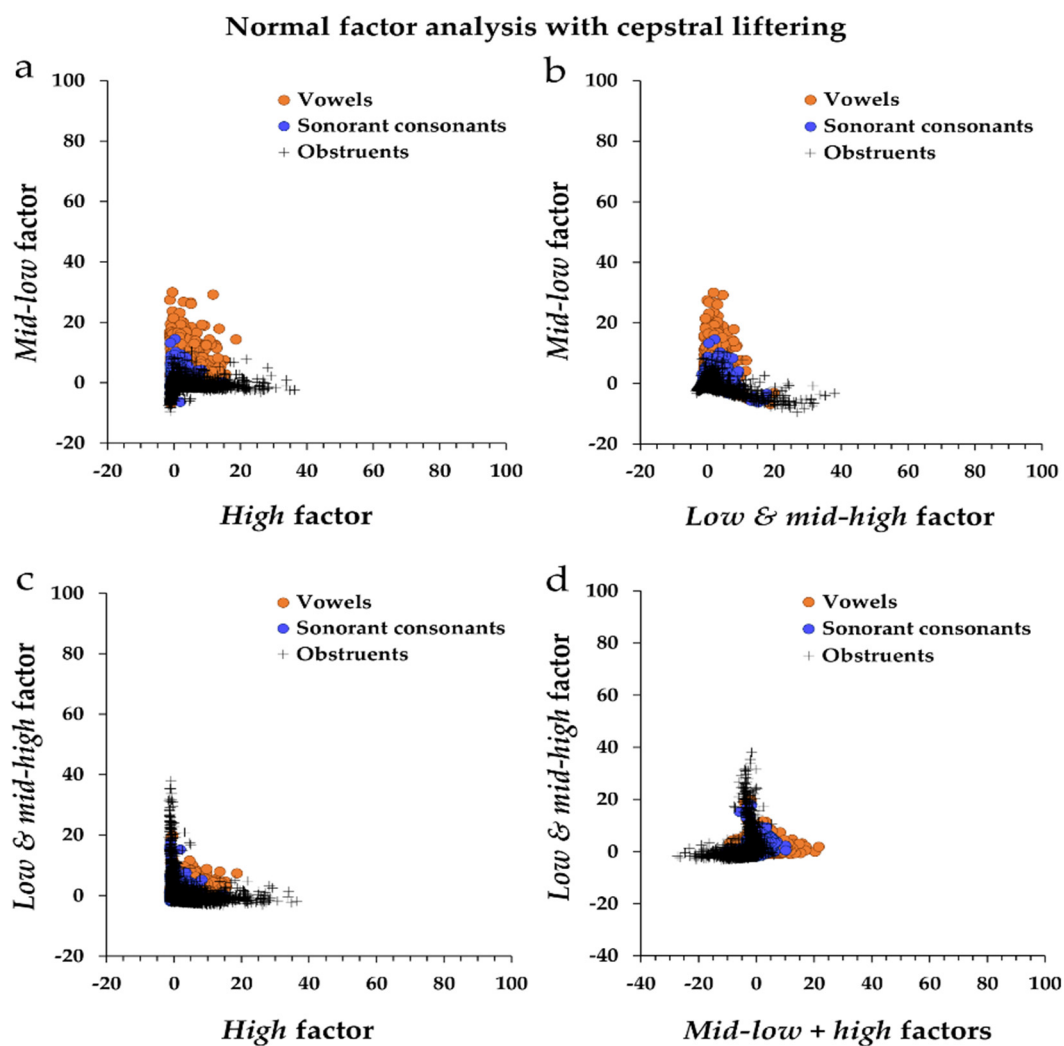


Figure 3. The distribution of factor scores by normal factor analysis with cepstral liftering. (a–c) represent all possible combinations of two factors out of the three factors, and (d) shows a two-dimensional mapping of the three factors.

The distribution of the factor scores showed a similar tendency in each of the four factor analyses (Figures 2–5). Although the cepstral liftering had some influence on the factor loadings of the second peak of the bimodal “low & mid-high factor”, overall the factor scores showed very similar distributions in the factor space. The same can be said for the use of the origin shift. Whether the origin shift was applied or not, the factor loadings (Figure 1a–d) and the factor scores (Figures 2–5) showed very similar tendencies.

As can be seen, all phonemes were distributed into three fairly distinctive areas in the three-dimensional space. On the “mid-low factor”, the highest factor scores were obtained by vowels, then by sonorant consonants, while the lowest factor scores were obtained by obstruents. Most of the obstruents occupied a position very near to or below zero on the “mid-low factor”. By contrast, they occupied a position above zero on the “low & mid-high factor”, while on the “high factor” they even reached the highest factor scores. In the three-factor factor analysis, Nakajima et al. (2017) [8] showed that the distribution of the three English-phoneme categories (i.e., vowels, sonorant consonants, and obstruents) constituted an L-shape in the two-dimensional factor space of the “high factor” and the “mid-low factor”. That is, the distributions of the phonemes linearly extended along both axes from the origin into the positive directions. Although not as clearly pronounced and narrowly shaped as in Nakajima et al. (2017) [8], also in the present results, a rather similar L-shaped distribution was

found as in Figures 2, 3, 4 and 5a. Interestingly, the distribution of the obstruents also showed an L-shape in the two-dimensional factor space of Figures 2, 3, 4 and 5c. The factor scores of obstruents were mutually exclusive between the “low & mid-high factor” and the “high factor”: If the factor score of the “low & mid-high factor” was higher, then the factor score of the “high factor” was very low, and vice versa. Generally, the L-shaped distribution of the factor scores of the obstruents across the “low & mid-high factor” and the “high factor” as analyzed here also appeared in Nakajima et al. (2017) [8], but hardly in a noticeable way. In the present analyses, the L-shaped distribution appeared far more prominently.

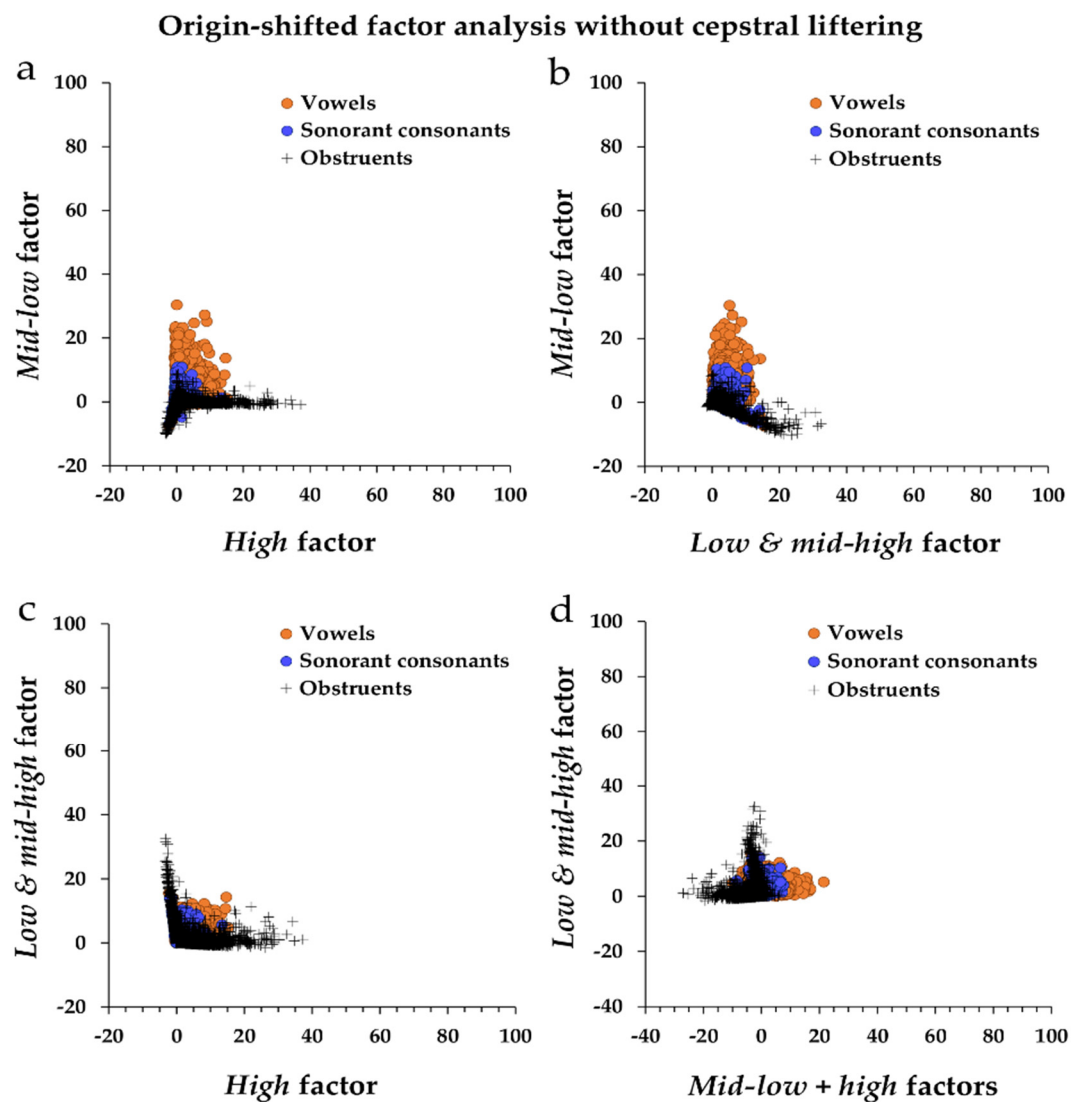


Figure 4. The distribution of factor scores by origin-shifted factor analysis without cepstral liftering. (a–c) represent all possible combinations of two factors out of the three factors, and (d) shows a two-dimensional mapping of the three factors.

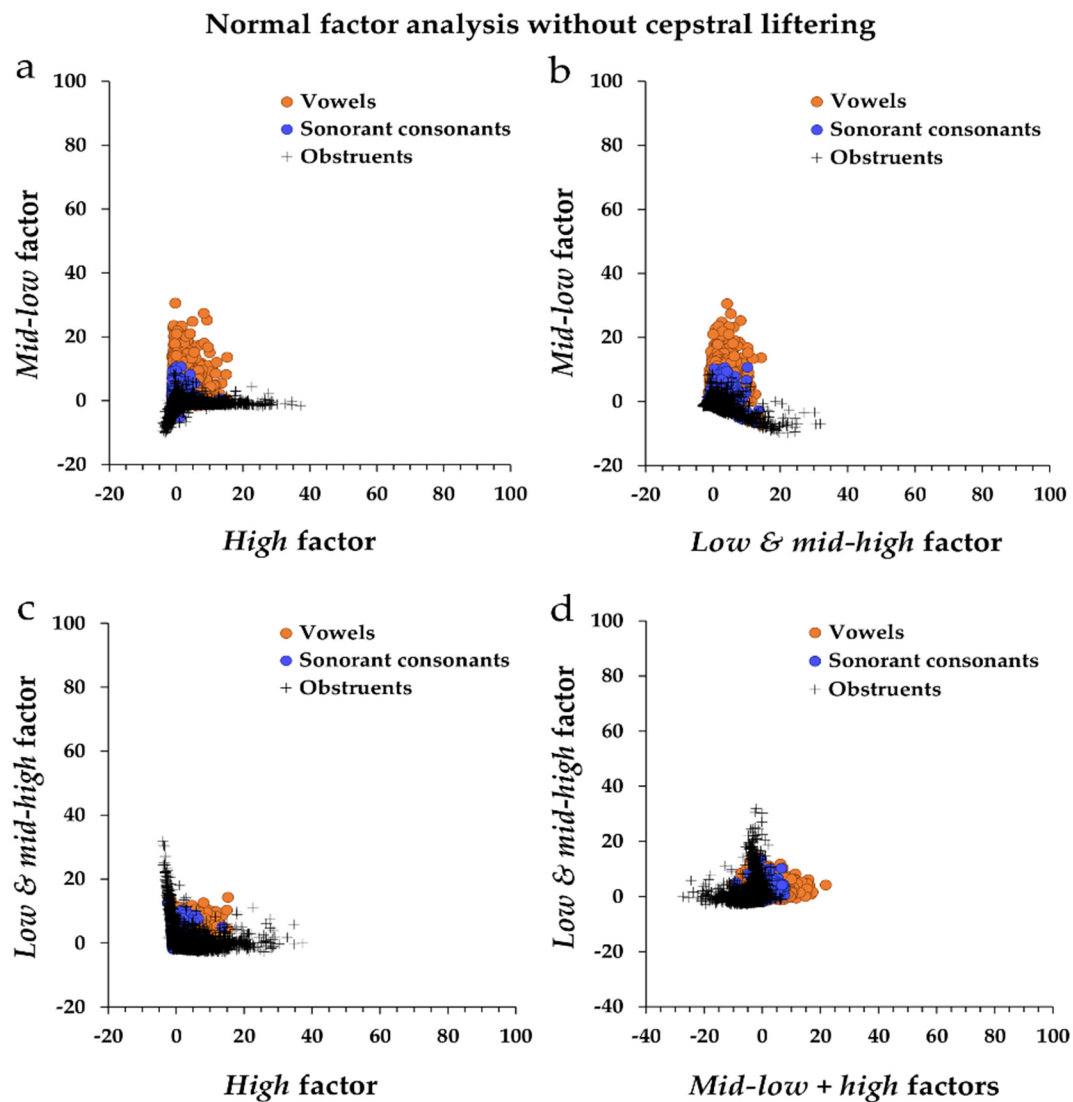


Figure 5. The distribution of factor scores by normal factor analysis without cepstral liftering. (a–c) represent all possible combinations of two factors out of the three factors, and (d) shows a two-dimensional mapping of the three factors.

Given the prominence of the L-shape for obstruents, more detailed analyses were performed. Figure 6 shows that the distributions of voiced obstruents (Figure 6a) and voiceless obstruents (Figure 6b) indeed showed distinctive L-shapes as well. Here eight obstruents were categorized as voiced obstruents, with the number of analyzed data points in parentheses: /b/ (609), /d/ (838), /g/ (274), /v/ (467), /ð/ (968), /dʒ/ (53), /ʒ/ (54) and /z/ (827). Nine obstruents were categorized as voiceless obstruents: /p/ (563), /t/ (1682), /k/ (786), /f/ (226), /h/ (340), /tʃ/ (195), /s/ (1389), /ʃ/ (552), and /θ/ (205). For the voiced obstruents (Figure 6a), three obstruents occupied a relatively high position on the “low & mid-high factor”, almost parallel to the y-axis, and close to the origin on the high factor: /b/, /v/, and /ð/, for which the factor scores of /v/ and /ð/ were around half of /b/. By contrast, four other voiced obstruents occupied a relatively high position on the “high factor”, almost parallel to the x-axis, and close to the origin on the “low & mid-high factor”: /g/, /ʒ/, /dʒ/ and /z/, for which the factor scores of /g/ and /z/ were around half of /ʒ/ and /dʒ/. Only the factor scores of /d/ were both distributed on the “low & mid-high factor” and the “high factor”, yet again in an L-shape.

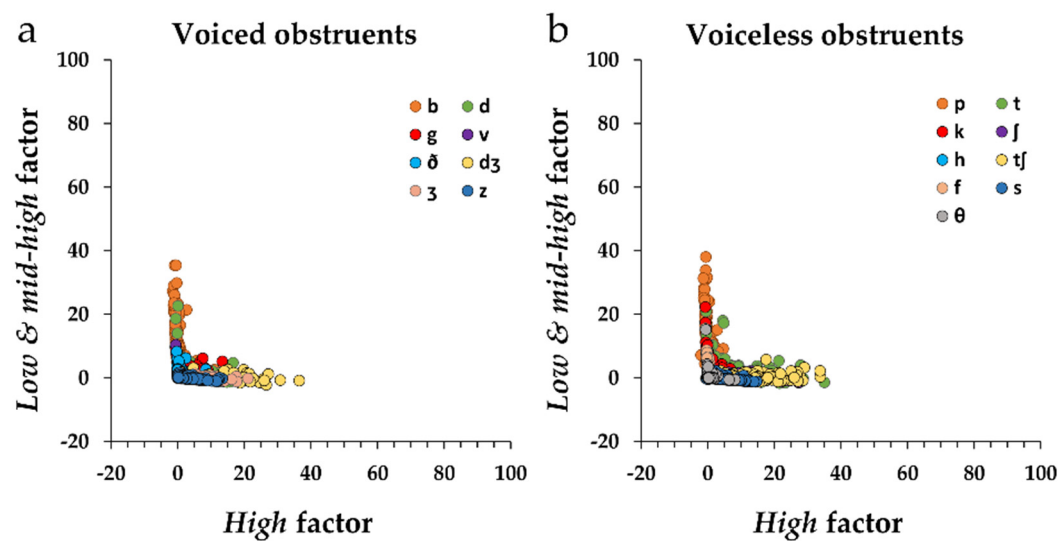


Figure 6. The L-shaped distributions of voiced obstruents (a) and voiceless obstruents (b) in the two-dimensional space of the “low & mid-high factor” and the “high factor”. Factor scores of the three extracted spectral factors were obtained by origin-shifted factor analysis with cepstral liftering (cf. Figure 2) and the English speech samples were from three native speakers.

The distributions of voiceless obstruents (Figure 6b), /p/, /k/, /h/, and /f/, displayed the L-shape, but they were mainly higher on the “low & mid-high factor”. The distributions of /t/ and /θ/ were also L-shaped, but they were mainly higher on the “high factor”. The distributions of three obstruents occupied a relatively high position only on the “high factor”, almost parallel to the x-axis, and close to the origin on the “low & mid-high factor”: /ʃ/, /tʃ/, and /s/, for which the factor scores of /s/ were about half of those of /ʃ/ and /tʃ/. In sum, the voiced obstruents /b/, /v/, and /ð/ were mainly related to the “low & mid-high factor”. The voiced obstruents /z/, /ʒ/ and /dʒ/ and the voiceless obstruents /ʃ/, /tʃ/, /s/ were mainly related to the “high factor”.

Figure 7 shows the distributions of all the obstruents divided into fricatives/affricates and plosives on the two-dimensional configuration, again clearly showing a distinctive L-shape. Here eleven obstruents were categorized as fricatives/affricates, with the number of analyzed data points in parentheses: /θ/ (205), /ð/ (968), /f/ (552), /v/ (467), /s/ (1389), /z/ (827), /ʃ/ (226), /ʒ/ (54), /h/ (340), /tʃ/ (195) and /dʒ/ (53). Six obstruents were categorized as plosives: /p/ (563), /t/ (1682), /k/ (786), /b/ (609), /d/ (838) and /g/ (274). For the fricatives/affricates (Figure 7a), the distributions of six obstruents occupied a relatively high position on the “high factor”, almost parallel to the x-axis, and close to the origin on the “low & mid-high factor”: /s/, /z/, /ʃ/, /ʒ/, /tʃ/ and /dʒ/. The distributions of five obstruents were both located on the “low & mid-high factor” and the “high factor” in an L-shape: /θ/, /ð/, /f/, /v/, /h/, but they were mainly higher on the “low & mid-high factor”. As for the distributions of plosives (Figure 7b), all of them displayed the L-shape, but /p/, /b/ and /k/ were mainly higher on the “low & mid-high factor”, and /t/, /d/ and /g/ were mainly higher on the “high factor”.

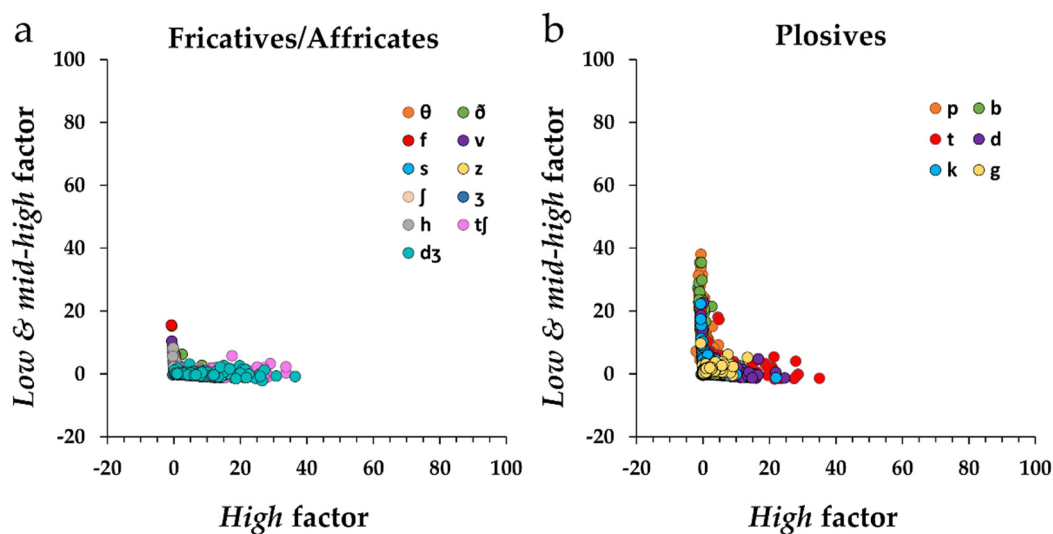


Figure 7. The L-shaped distributions of obstruents divided into fricatives/affricates (a) and plosives (b) in the two-dimensional space of the “low & mid-high factor” and the “high factor”. Factor scores of the three extracted spectral factors were obtained by origin-shifted factor analysis with cepstral liftering (cf. Figure 2) and the English speech samples were from three native speakers.

4. Discussion and Conclusions

Here we performed a comparative study between different factor analysis methods applied to English speech. Our first aim was to determine whether the use of normal factor analysis and a modified “origin-shifted” factor analysis caused differences in the extracted spectral factors when applied to English speech. The difference between these two methods was manifested in two processing aspects: (1) whether cepstral liftering was applied or not, (2) whether the origin shift was used or not. The present results showed that without cepstral liftering, the factor loadings of one factor were more pronounced (Figure 1), but neither cepstral liftering nor the origin shift had a large impact on the factor scores (Figures 2–5), whose distributions were similar.

The merit of the origin-shifted analysis is the following. Utilizing normal factor analysis, it can be difficult to find any features around silent parts in speech. When the speech is resynthesized, noise is very likely generated also at the silent parts, resulting in a continuous noise sounding in the background. The biggest advantage of the origin shift is that it makes all silent parts in speech signals plotted onto the silent point in the factor space, and that the silent parts remain silent when the speech is resynthesized. If the quality of the resynthesized speech is better, it can very likely be related more closely to the real auditory signal, which makes it more useful for listening experiments. Given that the results of the normal factor analysis and the origin-shifted factor analysis were similar, the origin-shifted factor analysis is highly recommendable for future research on speech resynthesis and subsequent listening experiments.

The second aim of our study was to determine the acoustic correlates of English phonemes by the four factor analyses. New insight was obtained with regard to obstruents. In an English syllable, an isolated obstruent always has a position as a syllable-onset or a syllable-end. If two or more obstruents are next to one another, one of them begins or ends the syllable. Analyzing obstruents acoustically as in the present study helped to understand this phonological phenomenon. Our analyses showed a factor with a frequency range around 1100 Hz. According to Nakajima et al. (2017) [8], this factor was only related to vowels and sonorant consonants [8]; it seems to be closely related to syllable nuclei, most of which are vowels but a few are sonorant consonants. We provided evidence that obstruents were not only associated with the factor related to a frequency range above 3300 Hz (the “high factor”), as suggested by Nakajima et al. [8], but also with the bimodal factor with frequencies around 300 Hz

and 2300 Hz (the “low & mid-high factor”). One likely cause for this is the difference in the initial analysis processing of the speech signal, but further investigation is required.

Our present findings thus suggest that these two extracted factors of the acoustic natures of obstruents reflect strong cues as to the beginning or the end of a syllable. It is important to note that the distributions of subsets of obstruents (voiced and voiceless, Figure 6; fricatives/affricates and plosives, Figure 7) indeed occupied high positions on the “low & mid-high factor” and the “high factor”, but not on the “mid-low factor”. The obstruents were often far from the origin, the silent point in the origin-shifted factor analysis, but they never went into the positive direction of the “mid-low factor”. This confirms that obstruents do not constitute the syllable nucleus, but rather delimit the syllable, corroborating the typical sonority hierarchy in phonology on which obstruents have the lowest position [10,11].

Further research is necessary to identify the acoustic correlates of individual obstruents in more detail. The distributions of most obstruents obtained a relatively high position on one of the two factors other than the “mid-low factor” (Figure 6, Figure 7). This should be investigated further in the future. It would be fruitful to identify acoustic correlates of consonant clusters. Consonant clusters in English are in many cases word-initial consonant clusters, such as /br/ in “bridge”, or word-final consonant clusters, such as /sk/ in “desks”. Based on purely acoustic analyses of English speech sounds as in the present study, we expect to perform listening experiments on consonant perception. Obstruents are very likely to be perceived as delimiting syllables by their lack of sonority, i.e., by their low score of the “mid-low factor”. An aspect of English phonology was thus connected to the acoustic natures of speech sounds.

Author Contributions: Conceptualization, Y.Z., Y.N., K.U. and G.B.R.; Methodology, Y.Z., Y.N. and G.B.R.; Software, T.K., Y.N. and Y.Z.; Formal Analysis, Y.Z., Y.N.; Investigation, Y.Z., Y.N. and G.B.R.; Resources, Y.N., K.U. and Y.Z.; Writing—Original Draft Preparation, Y.Z. and G.B.R.; Writing—Review and Editing, Y.Z., Y.N., K.U., T.K. and G.B.R.; Supervision, Y.N. and G.B.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC were funded by JSPS KAKENHI, grant numbers JP17H06197 and JP19H00630, and JST SCORE (to Y.N. in FY2019).

Acknowledgments: Y.Z. is supported by a MEXT Super Global University Project scholarship from the Japanese Ministry of Education, Culture, Sports, Science and Technology. We wish to thank Xiaoyang Yu for checking the samples from “The ATR British English Speech Database”.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. Zwicker, E. Subdivision of the audible frequency range into critical bands. *J. Acoust. Soc. Am.* **1961**, *33*, 248. [CrossRef]
2. Smith, J.O.; Abel, J.S. Bark and ERB bilinear transforms. *IEEE Trans. Audio Speech Lang. Process* **1999**, *7*, 697–708. [CrossRef]
3. Nakajima, Y.; Ueda, K.; Remijn, G.B.; Yamashita, Y.; Kishida, T. How sonority appears in speech analyses. *Acoust. Sci. Tech.* **2018**, *39*, 3. [CrossRef]
4. Plomp, R.; Pols, L.C.W.; van de Geer, J.P. Dimensional analysis of vowel spectra. *J. Acoust. Soc. Am.* **1967**, *41*, 707–712. [CrossRef]
5. Pols, L.C.W.; Tromp, H.R.C.; Plomp, R. Frequency analysis of Dutch vowels from 50 male speakers. *J. Acoust. Soc. Am.* **1973**, *53*, 1093–1101. [CrossRef] [PubMed]
6. Zahorian, S.A.; Rothenberg, M. Principal-components analysis for low-redundancy encoding of speech spectra. *J. Acoust. Soc. Am.* **1981**, *69*, 832–845. [CrossRef]
7. Ueda, K.; Nakajima, Y. An acoustic key to eight languages/dialects: Factor analyses of critical-band-filtered speech. *Sci. Rep.* **2017**, *7*, 42468. [CrossRef] [PubMed]
8. Nakajima, Y.; Ueda, K.; Fujimaru, S.; Motomura, H.; Ohsaka, Y. English phonology and an acoustic language universal. *Sci. Rep.* **2017**, *7*, 46049. [CrossRef] [PubMed]
9. Féry, C.; Vijver, R. *The Syllable in Optimality Theory*; Cambridge University Press: Cambridge, UK, 2003; p. 356.

10. Spencer, A. *Phonology: Theory and Description*; Blackwell Press: Oxford, UK, 1996.
11. Harris, J. *English Sound Structure*; Blackwell Press: Oxford, UK, 1994; pp. 47–56.
12. Kishida, T.; Nakajima, Y.; Ueda, K.; Remijn, G.B. Three factors are critical in order to synthesize intelligible noise-vocoded Japanese speech. *Front. Psychol.* **2016**, *7*, 517. [[CrossRef](#)] [[PubMed](#)]
13. Campbell, N. ATR Interpreting Telephony Research Laboratories. In *The ATR British English Speech Database*; ATR: Kyoto, Japan, 1993; TR-I-0363.
14. Lennig, M.; Brassard, J.P. Machine-readable phonetic alphabet for English and French. *Speech Commun.* **1984**, *3*, 165–166. [[CrossRef](#)]
15. Nyquist, H. Certain topics in telegraph transmission theory. *AIEE Trans.* **1928**, *47*, 617–644. [[CrossRef](#)]
16. Zwicker, E.; Terhardt, E. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.* **1980**, *68*, 1523–1525. [[CrossRef](#)]
17. Awan, N.S.; Giovinco, A.; Owens, J. Effects of vocal intensity and vowel type on cepstral analysis of voice. *J. Voice* **2012**, *26*, 670.e15–670.e20. [[CrossRef](#)] [[PubMed](#)]
18. Kaiser, H.F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **1958**, *23*, 187–200. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).