Journal of Otolaryngology and Neurotology Research

JONR, 5(1): 112-115
www.scitcentral.com

ISSN: 2641-6956

Original Research Article: Open Access

The Effect of Coarticulation on Speaker Identification in Disguised Voices

Lakshmi Prasanna P[*]

[*]*Associate Professor in Speech-Language Pathology, Helen Keller's Institute of Research and Rehabilitation for the Disabled Children, RK Puram, Secunderabad-06, Telangana, India.*

Received October 21, 2022; Revised December 30, 2022; Accepted January 02, 2023

ABSTRACT

The present study was done to identify a speaker from the disguised voices by using voice changing app. One female speaker aged 35 years was participated in the present study. Participant was asked to produce two sentences which were later converted into different disguise voice by using voice changer app, which were audio recorded and transferred to the system for further analysis. Formants of vowel /i: / were analyzed by using PRAAT software. The results indicated that vowel /i:/ showed better speaker identification when it is coarticulated with a consonant /t/ than /k/. When PRAAT original compared with APP original along with Man and Baby type voices, PRAAT original and APP original are identified best than other voices in voice changer APP. Finally author concluded that SPID can be better when compared with APP voice with Lab (PRAAT) voice than other conditions.

Keywords: Formants, Voice changer app, Disguised voice, Speaker identification, PRAAT

INTRODUCTION

Speaker recognition can be divided into speaker identification and speaker verification. Speaker identification can be done through close set or open set identification. Speaker recognition can again be divided into naïve and technical recognition.

There are three methods of speaker identification by listening/perceptual method, by visual method and by machine. There have been several measures for speaker identification, first and second formant frequencies [1], formants and Bandwidths, higher formants, fundamental frequency, pitch contour, Linear Prediction Coefficients, Cepstral Coefficients & Mel Frequency Cepstral Coefficients, Long Term Average Spectrum and Cepstrum, vowel space in disguise have been used in the past.

The frequency peaks of a speech are formants which are also known as resonant frequencies, appear in a speech stream. According to the speaker's vocal tract, these frequencies resonate. The frequencies connected to vowel sounds in a language are referred to as vowel formants. It is commonly established that, in most situations, the two or three lowest vowel formants are adequate to discriminate between vowels, Aalto [2].

A predefined approach is used by speakers to change their regular articulator movements, such as tongue positioning, during voice disguise.

Su, Li & Fu [3] done a study by nasal spectra, the study was focused on the changes of coarticulation of vowels with nasal sounds. The spectral differences between the mean spectra of nasals followed by front vowels and those of nasals followed by back vowels were used as the acoustic measure of the coarticulation of [m] and [n] with the following vowel [V]. Authors stated that coarticulation was found to give more reliable clues than the nasal spectrum alone, which had earlier been found to be one of the best acoustic clues for identifying speakers.

Paliwal [14] used eleven English vowels (/ɜ/, /ʌ/, /u/, /ʊ/, /o/, /ɔ/, /a/, /ae/, /ɛ/, /i/, /i/) are tested in an experiment to see which ones are most effective at identifying speakers automatically. For speaker identification, a minimal distance classifier is utilized together with the first four formant frequencies of the vowel sounds as parameters. The best

Corresponding author: Lakshmi Prasanna P, Associate Professor in Speech-Language Pathology, Helen Keller's Institute of Research and Rehabilitation for the Disabled Children, RK Puram, Secunderabad-06, Telangana, India, Tel: +91 8977034115; E-mail: lakshmiprasannaspeechpathologist@yahoo.com

Citation: Prasanna LP. (2023) The Effect of Coarticulation on Speaker Identification in Disguised Voices. J Otolaryngol Neurotol Res, 4(1): 112-115.

performance for speaker identification was determined to be with the vowel. Lakshmi Prasanna and Savithri [1] had done a study on formants in various vowels such as /a/ /i/ and /u/ which were embedded in words. The results showed that /i:/ has benchmarking of 75% (better) compared with vowels /a:/ and /u:/. Vowel /u:/ had poor benchmarking. Authors stated that vowel /i:/ had 75% benchmarking $F_2 \approx F_1$ of vowel /i:/ could be used for speaker identification.

Lakshmi Prasanna and Savithri (2011) had done a study on benchmarks for SPID in various nasal continuants in Telugu language by using formants and bandwidths which reveals Nasal /ṇ/ showed 80% correct identification when followed by a vowel /i/ and 70% for the nasal /ṉ/when followed by vowel /u/ in females. Group C; in males, nasal /ñ/ had 90% correct identification and /ṇ/ had 70% correct identification. In females /ṇ/ had 80% correct identification and /ŋ/ and /ṉ/ had 70% correct identification. Palatal nasal continuant /ñ/ had the highest percent correct identification. Of the vowel contexts, context of /i/ in group C was the best with 90% correct identification.

Tavi [5] had done a study on articulation during voice disguise. The authors examined the use of foreign accent imitation as a voice disguise strategy and developed a novel way to uncover articulatory differences between natural speech and voice disguise, functional t-tests. Identified articulatory disparities were also assessed in relation to the effectiveness of an x-vector-based automatic speaker verification system.

The first two formants that demonstrated a significant change in vowel space across speaking speeds in a consonant-specific way were employed in a study by Narayanan [6] on an auditory articulatory database of VCV sequences and words in Toda. Results showed that most VCV sequences exhibit a significant effect of consonant context on both anticipatory and carryover articulation. Coarticulation significantly decreases when speaking more quickly. The study shows that there are variations in how rate and consonant context impact coarticulatory structure.

The review indicates that earlier studies used aural and visual methods for SPID using formants, power spectra and Perceptual Linear Predictive (PLP). There has been no particular research done till date using coarticulation effect on the formants to identify a speaker. In this context the present study was planned to check “the impact of coarticulation on speaker identification among voice disguises”.

METHOD

Subjects: A 35-year-old female took part in the study. I want to Tweet her and I want to Tweek her are two statements made up of the two words "Tweek" and "Tweet", respectively. Except the final phoneme, the rest of the words' word structures are identical for both. This word structure was designed to determine whether coarticulation varies due to different phones in similar word structures. The participants were told to casually repeat the statements three times.

Analysis: These were recorded by using a “Voice Changer App” in Samsung A30s smart phone. Various disguise voices were selected such as Man, Baby, App Original and voice recording was also done by using PRAAT software. The obtained data was converted “OGG” file into “WAV” file by using online portal to avail easily in PRAAT software. The data was transferred on to the computer memory. Wide band spectrograms of the words were displayed using the analysis program of PRAAT; 6.2.13 version [7]. Praat Original voices consider as unknown and the other voices such as App Original, Man and Baby considered as known speakers (disguise voices). Formants (F1 and F2) of vowel /i:/ were measured for all the speakers in both the words. Closed-set speaker identification tasks were performed, in which the examiner was aware that the “unknown” speaker was among the “known” ones.

RESULTS

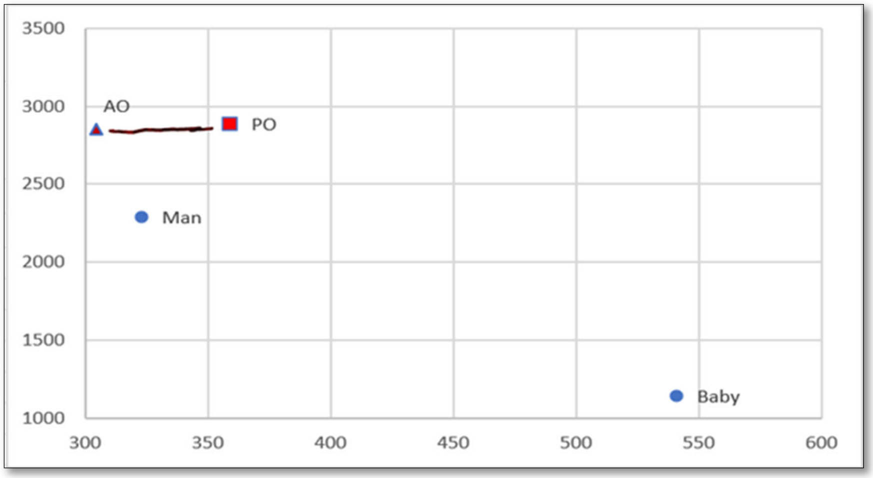
In all the voices, the formant frequencies (F1 and F2) for both words were measured. The word "Tweek" contains greater formants than "Tweet," which indicates that formant frequencies are altered as a result of the co-articulation of phonemes. When compared to other voice types, the "Baby" voice type showed high F1 and lower F2 formants, whereas the "Man" voice type displayed lower F1 and F2 formants. Other than the F1 of the App Original and Baby type voices, the word "Tweek" had higher F1 and F2 formants. This might be because of the coarticulatory influence of the voiceless stop /k/ (velar) on the /t/ sound (alveolar, voiceless stop). Both the constriction for the phonemes "k" and "t" differs. The information was included in **Table 1**.

Table 1. Shows formant frequencies (F1 & F2) of words “Tweek” and “Tweet” in all voices.

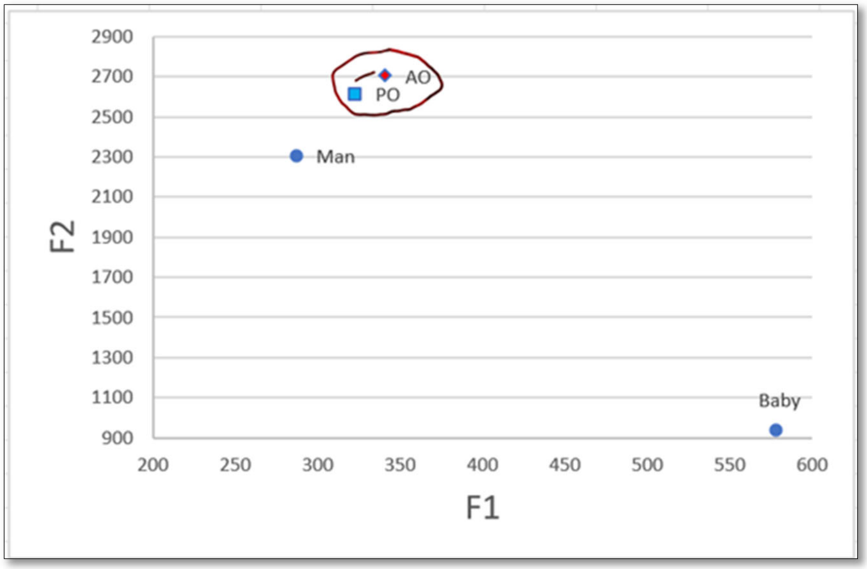
Types of voice	Tweek Formants (Hz)		Tweet Formants (Hz)	
	F1	**F2**	**F1**	**F2**
Praat Original	359	2887	323	2614
App Original	305	2851	341	2705
Baby	541	1142	578	941
Man	323	2287	287	2305

Speaker identification (SPID): Tweek vs. Tweet

The coarticulation effect was seen in the process of a closed set of speaker identification where it was very different in two similar words with a minimal change. Praat Original (PO) was considered as unknown speaker and other disguise voices such as Man and Baby type along with a direct voice recorded in the APP i.e., APP original (AO) was considered as known speakers. Interestingly, the voice recorded in the PRAAT (unknown) as well as App original (known) were near rather than disguise type of voices in the word "Tweek". The data was depicted in the **Graph 1**. In the other hand, word "Tweet" showed a very closed speaker identification on unknown speaker (PO) with a known speaker (AO), whereas the other two are not identified. The data was plotted in the **Graph 2** where F1 on X axis and F2 on Y axis were considered.



Graph 1. Shows SPID among closed set of speakers for a word "TWEEK" AO: App original, PO: PRAAT original.



Graph 2. Shows correct SPID among closed set of speakers for a word "TWEET" AO: App original, PO: PRAAT original.

DISCUSSION

Glenn & Kleiner [8] derived a benchmark of 43% for 300 vectors, 62 % for 150 vectors and 82 % for 60 vectors using power spectra of nasal phonation /n/. However, in the present study palatal nasal /ñ/ had the best correct percent identification in the context of vowel /i/. The results of the present study are also not in consonance with those of Ying-Yong Qi & Robert (1992) who used perceptual linear predictive method (PLP) for the initial nasal consonants /m/ and /n/ and derived 83 % of benchmark. Lakshmi Prasanna and Savithri [1] used formants and stated that vowel /i:/ can be used for SPID. Jakhar [9] and Srividya [10] used cepstra of vowels and reported a benchmarking of more than 80%. Lakshmi Prasanna and Savithri (2011) used bandwidths and formants of nasal continuants and stated that the results suggest that the benchmarking will be best in the context of vowel /i/ for the palatal nasal / ñ/. Hence, it is recommended that the palatal nasal /ñ/ be used for speaker identification. In general, it could be concluded that nasals /ñ/ and / ṇ / in the context of vowels /a/ and /i/ could be used for speaker identification as their benchmarking is ≥ 70% when two speakers are compared. Based on the previous research the present study author considered vowel/i: with two different consonants to check the coarticulation which reveals the co articulation effect has a role in SPID where the results showed greater SPID was found with word "Tweet" than a word "Tweek".

CONCLUSION

Finally, the author of the present study concluded that coarticulation effect should also be taken into account in the SPID, where the alterations were noticeable in two different phonemes articulated with the same vowel. Technology is one of the aspects of modern world that is developing the fastest, with new strategies emerging daily, a wide range of mobile devices and features, and an easy availability of free APPS for download. Daily increases in crime are also being seen. To uncover a specific characteristic that can be used to identify a speaker in this situation, numerous investigations should be conducted. Therefore, numerous investigations should be conducted in this context to determine a specific attribute that can be used to identify a speaker. In contrast to other disguised voices, APP original recordings may be compared to lab recordings of a specific person. The coarticulatory influence of the voiceless stops /k/ (velar) and /t/ (alveolar, voiceless stop), which are caused by different constrictions in the vocal tract, may be the reason why the vowel /i/ is more easily recognized when it is coarticulated with a consonant /t/ than /k/. The current study only includes one participant, Indian English language and a small number of words that begin with the vowel /i/. The findings can be applied to large set of data, other vowels, different languages, telephone recording situations, and disguise settings. Future research is necessary in these areas.

CONFLICT OF INTEREST

The author declared that there are no conflicts of interest.

ACKNOWLEDGEMENT

Sincere thanks to the participant and Helen Keller Institute for allowing the author to complete the study.

SOURCE OF FUNDING

There is no funding for this study and was done under the part of Research at Helen Kellers Institute, Secunderabad, India.

REFERENCES

1. Lakshmi P, Savithri SR (2009) Benchmark for speaker identification using vector $F_2 \approx F_1$ vector, FRSM-09 Proceedings.
2. Aalto D, Huhtala A, Kivela A, Malinen J, Palo P, et al. (2012) Vowel formants compared with resonances of the vocal tract.
3. Su L, Li K-P, Fu KS (1974) Identification of speakers by use of nasal coarticulation. J Acoust Soc Am 56(6): 1876-1883.
4. Paliwal KK (1984) Effectiveness of different vowel sounds in automatic speaker identification. *J Phonetics 12*: 17-21.
5. Tavi L, Kinnunen T, Meister E, González-Hautamäki R, Malmi A (2021). Articulation during Voice Disguise: A Pilot Study. In: Karpov, A., Potapova, R. (eds) Speech and Computer. SPECOM 2021. Lecture Notes in Computer Science, Vol: 12997. Springer.
6. Narayanan S, Illa A, Anand N, Sinisetty G, Narayanan K, et al. (2019). An acoustic-articulatory database of VCV sequences and words in Toda at different speaking rates. 2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), pp: 1-6.
7. Boersma P, Weenink D (2009) PRAAT 5.1.14 software.
8. Glenn JW, Kleiner N (1968) Speaker identification based on nasal phonation. J Acoust Soc Amer 43(2): 368-372.
9. Jakhar SS (2009) Benchmarks for speaker identification using Cepstral coefficient in Hindi. Unpublished project of Post graduate Diploma in Forensic Speech Science and Technology submitted to university of Mysore, Mysore.
10. Srividya C (2010) Benchmarks for speaker identification using Cepstral coefficient in Kannada. Unpublished project of Post graduate Diploma in Forensic Speech Science and Technology submitted to university of Mysore, Mysore.