Highly spectrally undersampled vowels can be classified by machines without supervision

Kathiresan, Thayabaran ; Maurer, Dieter ; Dellwo, Volker

DOI: https://doi.org/10.1121/1.5111154

Highly spectrally undersampled vowels can be classified by machines without supervision

Thayabaran Kathiresan, Dieter Maurer, and Volker Dellwo

Citation: The Journal of the Acoustical Society of America **146**, EL1 (2019); doi: 10.1121/1.5111154
View online: https://doi.org/10.1121/1.5111154
View Table of Contents: https://asa.scitation.org/toc/jas/146/1
Published by the Acoustical Society of America

Highly spectrally undersampled vowels can be classified by machines without supervision

Thayabaran Kathiresan
Institute of Computational Linguistics, University of Zurich, Andreasstrasse 15, CH-8050, Zurich, Switzerland
thayabaran.kathiresan@uzh.ch

Dieter Maurer
Institute for the Performing Arts and Film, Zurich University of the Arts, Pfingstweidstrasse 96, CH-8031, Zurich, Switzerland
dieter.maurer@zhdk.ch

Volker Dellwo[a)]
Institute of Computational Linguistics, University of Zurich, Andreasstrasse 15, CH-8050, Zurich, Switzerland
volker.dellwo@uzh.ch

Abstract: An unsupervised automatic clustering algorithm (k-means) classified 1282 Mel frequency cepstral coefficient (MFCC) representations of isolated steady-state vowel utterances from eight standard German vowel categories with f_o between 196 and 698 Hz. Experiment I obtained the number of MFCCs (1–20) in connection with the spectral bandwidth (2–20 kHz) at which performance peaked (five MFCCs at 4 kHz). In experiment II, classification performance with different ranges of f_o revealed that ranges with $f_o > 500$ Hz reduced classification performance but it remained well above chance. This shows that isolated steady state vowels with strongly undersampled spectra contain sufficient acoustic information to be classified automatically.

[BHS]
Date Received: March 7, 2019 **Date Accepted:** May 14, 2019

1. Introduction

Fundamental frequency of oscillation (f_o) in human voice can vary with respect to the audience or situation. Hess (1983) reviews studies showing that vocal ranges can substantially exceed 1 kHz (with extreme upper limits above 2 kHz) depending on production type (above all phonation type) and gender and age group. Maurer (2016) documents that the upper f_o range in speech of untrained female speakers as well as journalists, TV hosts, and actresses is approximately 900 Hz (excluding higher f_o levels for exclamations).

Studying vowels at high f_o is problematic. As f_o increases, the spacing between harmonics increases, resulting in a less dense sampling of the vocal tract transfer function [henceforth: spectral undersampling (Goldstein, 1980)]. This leads to a highly sparse acoustic representation of formant frequencies which are often believed one of the most crucial acoustic correlates in vowels. The phenomenon is aggravated when f_o exceeds the first vocal tract resonance (Sundberg, 2012) which—as a rule of thumb—starts showing effects at an f_o of about 300 Hz (Ferreira, 2007). This led to the general view that vowels become more unintelligible with increasing f_o and that they fully lose the category information at $f_o \sim 500$ Hz (Sundberg, 2012). The view is supported (a) by findings from classical European singing styles, revealing that all vowels higher than 500 Hz are heavily biased towards the perception of /a/ [Sundberg (2012); review in Friedrichs *et al.* (2015); Maurer (2016), pp. 35–37, 107–111] and (b) by the fact that formant estimation algorithms cannot provide meaningful formant estimates when $f_o > 300$ Hz (Ferreira, 2007).

Despite the findings from some studies suggesting that vowel quality degrades with increased undersampling, it is surprising that other studies found that humans show high performance at recognizing naturally produced vowels when spectral undersampling is high and $f_o >$ F1. Smith and Scott (1980) showed that front vowels were highly intelligible up to 880 Hz when produced in isolation with a raised larynx and argued that a possible loss of intelligibility at this f_o is related to western singing styles. Such styles maintain esthetic characteristics and strength against orchestras over intelligibility (Joliveau *et al.*, 2004). This

[a)]Author to whom correspondence should be addressed.

argument is supported by Maurer and Landis (1995) who showed that listeners can identify vowels produced in isolation by untrained men, women and children up to $f_o = 850$ Hz and by Maurer *et al.* (2014) who showed that listeners can identify vowels in singing styles like Cantonese opera up to $f_o \sim 700$ Hz. Friedrichs *et al.* (2015) carried out word identification in minimal pairs where the contrast was vocalic. They found a ceiling performance with f_o as high as 880 Hz when vowels were in word context and close to ceiling performance when consonantal environment and coarticulatory information were removed. This reveals that the linguistic function of vowels is still maintained their steady-state parts at f_o higher than typical for conversational speech. Friedrichs *et al.* (2017) increased the difficulty of the task and showed that vowel identification performance in multiple choice (eight standard-German vowels) is high until $f_o \sim 880$ Hz. Performance for the three corner vowels /a/, /i/, /u/ remains high until 1046 Hz while performance for non-corner vowels sometimes dropped to chance. All studies (Smith and Scott, 1980; Maurer *et al.*, 2014; Friedrichs *et al.*, 2015; Friedrichs *et al.*, 2017) argued that production styles heavily influence vowel recognizability and suggest that the spectral variability introduced by vocal tract undersampling and the lack of formant information only has a small effect on human vowel category perception. This supports views which argue that the role of formants in human vowel identification has probably been overestimated (Kiefte *et al.*, 2013) and that the relationship between source and filter in encoding vowels is complex, in particular when spectral undersampling occurs (Maurer and Landis, 1995).

Given that humans can reliably obtain cues to vowel identity from isolated steady-state vocalic intervals with f_o up to 880 Hz (Smith and Scott, 1980; Maurer and Landis, 1995; Maurer *et al.*, 2014; Friedrichs *et al.*, 2015; Friedrichs *et al.*, 2017), we were interested whether computers can perform similar tasks. This was tested in the present paper using Mel frequency cepstral coefficients (MFCCs) of vocalic utterances with various degrees of undersampling. To arrive at a static spectral representation of a vocalic utterance, we averaged MFCC frames from each vowel to arrive at one single MFCC representation for a vocalic utterance. These representations were classified with an unsupervised k-means clustering algorithm. MFCCs are one of the most widely used acoustic representations in contemporary speech technology (Weinstein *et al.*, 2016; Choi *et al.*, 2016). Previous studies revealed that MFCCs are highly suitable acoustic representations for automatic vowel recognition (de Wet *et al.*, 2004; Ferreira, 2007). Ferreira (2007) found that MFCCs are not influenced by f_o variability but this result was based on vowels with a maximum f_o of 400 Hz. An increased f_o will lead to more spectral undersampling, presumably leading to a poorer representation of vocal tract contributions in the output signal. In the present study we thus increased spectral undersampling by increasing f_o to 698 Hz (F4 note in musical scale).

Two parameter settings in obtaining MFCCs strongly influence machine recognition performance: (a) the number of coefficients (Zheng *et al.*, 2001) and (b) the overall spectral bandwidth of the signal. Previous studies are not consistent in the use of these parameters. Typically, the number of Mel filters in MFCC extraction procedure ranges from 20 to 40, followed by discrete cosine transform (DCT). Of these 20+ coefficients, only the first 12 to 15 are most typically applied in automatic speech or speaker recognition applications (de Wet *et al.*, 2004; Choi *et al.*, 2016). This is because the higher MFCC coefficients represent fast changes in the filterbank energies which are assumed to take a negative impact on performance. The first coefficient (MFCC 1) is the so-called "energy coefficient" as it represents filterbank energy (Zheng *et al.*, 2001), the next higher coefficients are numbered sequentially. To classify vocalic utterances, de Wet *et al.* (2004) used the first 13 MFCCs as well as 3 MFCCs without the energy coefficient (i.e., MFCC 2, 3, and 4) at 8 kHz spectrum. Ferreira (2007) used 16 MFCCs without energy coefficient, i.e., from MFCC 2 to 17 at 16 kHz spectrum. Zheng *et al.* (2001), however, argue that including the energy coefficient is always beneficial. The signal bandwidth in previous studies is typically 8 kHz (e.g., 16 kHz sampling frequency in de Wet *et al.*, 2004) while some studies used higher bandwidth 16 kHz (e.g., 32 kHz sampling frequency in Ferreira, 2007). It is thus difficult to compare performances in previous studies and—more importantly—it remains unclear which methodological choice (signal bandwidth in combination with MFCCs) leads to the highest vowel recognition performance.

In experiment I, we used an automatic unsupervised k-means classifier (Vallabha *et al.*, 2007) to test which number of MFCCs in connection with which signal bandwidth shows the best recognition performance in a vowel dataset with large f_o variability (between 196 and 698 Hz). We chose unsupervised classification as this technique is frequently used as a computational model to simulate human phoneme acquisition (Guevara-Rukoz *et al.*, 2017). This automatic processing is probably the closest

technical analogy to some human perception tasks in which vowels are classified into abstract units based on token similarities without prior linguistic knowledge of the particular language (e.g., in early language acquisition). Further, from a methodological point of view, we found that the unsupervised classification is the best choice considering the sample size (max N = 1282, see Sec. 2.1) which would be inappropriately small to train a supervised clustering algorithm. Unsupervised classification performance was measured by entropy (Tan *et al.,* 2005).

In experiment II, we used the setting that showed the best performance in experiment I and tested the effect of f_o on vowel recognition performance, by comparing performances for different data subsets in which f_o varies in range from a low (196 to 295 Hz) to a high range (587 to 698 Hz). In experiment II, we also compared the performance for the best parameter setting from experiment I with typical settings applied in previous research. Identical classifier and performance measure were used as in experiment I.

2. Experiment I: Obtaining the best MFCCs and bandwidth setting for vowel classification using a k-means unsupervised classifier

Vowel recognition performance was tested in 8 Standard German vowels on a dataset with high f_o variability between 196 and 698 Hz. We systematically varied the number of MFCCs (cumulative from 1 to 20; always including the energy coefficient; Zheng *et al.,* 2001) and spectral bandwidth (from 2 to 20 kHz). We expected that 1 MFCC should perform worst since energy alone should not be sufficient for vowel classification, even though it might show some performance since openness of vowels is to some degree revealed in their energy (energy in open vowels is higher than in closed ones). Performance was expected to be best around 13 MFCCs as this is the commonly used number (de Wet *et al.,* 2004; Ferreira, 2007). However, given that vocalic information is predominantly spread in lower coefficients after DCT a lower number might be feasible but is possibly variable with spectral bandwidth.

2.1 Method

2.1.1 Speakers and recording procedure

Four female professional stage actresses produced eight isolated steady-state Standard German vowels (/i/, /e/, /y/, /ø/, /ɛ/, /a/, /o/, /u/) at 14 different f_o levels ranging between 196 and 698 Hz (196, 220, 247, 262, 294, 330, 349, 392, 440, 494, 523, 587, 659, 698 Hz) and three different vocal efforts (low, medium and high); N = 1344 (4 speakers * 8 vowels * 14 f_o levels * 3 vocal efforts). All utterances were produced by the speakers in non-professional (non-style) productions, which means that the intelligibility of the vowels was favored over esthetics. The data were taken from a larger vowel database by Maurer *et al.* (2018). Speakers were presented a reference tone corresponding to the respective f_o level prior to production via loudspeaker. Speakers were recorded in standing position in a noise-controlled room with a speaker-microphone distance of 30 cm. The sounds were recorded on a PC (cardioid condenser microphone Sennheiser MKH 40 P48, pop shield, audio interface Fireface UCX; 16 bit, 44 100 samples/s, PCM). Five phonetic expert listeners (professionally trained singers and actors) did an identification task in a multiple-choice identification paradigm. We selected all vowels with a recognition rate of 3 votes or above out of 5 listeners (1028 tokens with 5/5 votes, 166 tokens with 4/5 votes, and 88 tokens with 3/5 votes) resulting in 1282 vowels. This ensured that the vowel category was typically identifiable for human listeners. Sixty-two vowel tokens had lower ratings (2 votes or less out of 5 listeners). These tokens were distributed rather equally across vowel categories and f_o levels (except levels 247, 262, and 523 Hz where no token was removed).

2.1.2 Feature extraction

MFCCs were extracted from central 0.3 s interval of every vowel recording using MATLAB R2018b (Young *et al.,* 2006). MFCC extraction parameters: Hamming window with frame duration (25 ms), frameshift (10 ms), pre-emphasis coefficient (0.95), number of filterbank channels (20), liftering parameter (22) were kept constant, N frames = 28. The number of MFCCs were varied between 1 to 20 (including MFCC 1; Zheng, 2001), spectral bandwidth varied from 2 to 20 kHz in steps of 1 kHz. The extracted array of MFCC features from all frames was averaged to result in a single MFCC representation per vowel recording.

2.1.3 Unsupervised clustering and entropy performance measure

The MFCC arrays of the 1282 vowel recordings were clustered into eight groups using a k-means vector quantization algorithm (Voicebox tool in MATLAB; Brookes, 2011). The initial parameters (mean of eight clusters) were set to be random and the

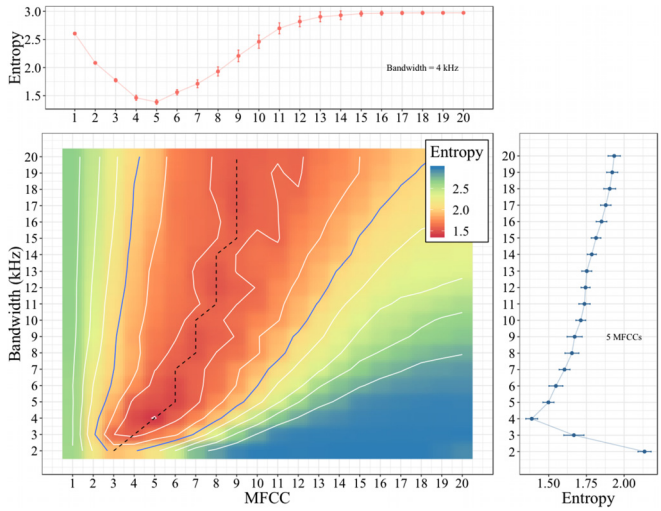

Fig. 1. (Color online) Large plot (bottom left) shows mean entropy of unsupervised vowel classification for bandwidth as a function of MFCC on a color-gradient scale (averaged over 500 experimental repetitions). f_o range was between 196 and 698 Hz. The dotted black line in the center-plot shows the minimum entropy values for the respective bandwidth-MFCC combination. The top- and right-plot shows the horizontal and vertical cross sections at 4 kHz bandwidth and at 5 MFCCs, respectively (dots are mean values and bars are standard deviation from 500 experimental repetitions).

algorithm was allowed to perform 100 iterations (maximum) to cluster the data. Entropy was used to evaluate the clusters as described by Tan *et al.* (2005). Here, the entropy scale ranges from 0 (highest entropy) to 3 (chance-level; calculated by 8 possible vowels: $-\log_2[1/8] = 3$). Matrix entropy is a continuous measure to obtain an unambiguous classification performance of an unsupervised classifier.

2.2 Results and discussion

Figure 1 (bottom left) shows mean entropy (averaged over 500 experimental repetitions) for unsupervised vowel classification with cumulative number of MFCCs (x axis), bandwidth (y axis), and entropy in color gradient (dark = high performance, and bright = low performance). The red-dotted line corresponds to the minimum entropy mean. The figure reveals that there is a clear performance peak (lowest entropy mean: 1.387) at five MFCCs in combination with a 4 kHz bandwidth. The cross-section of performance at 4 kHz bandwidth for all MFCCs (Fig. 1, top) and the cross-section of performance at five MFCCs at all bandwidths (Fig. 1, bottom right) are provided for better legibility of entropy numbers and for showing standard deviations of entropy means (over 500 experimental repetitions). Apart from the five MFCCs and 4 kHz bandwidth combination, high performance could be revealed (red line) with MFCCs ranging between 3 and 9 at bandwidths between 2 and 20 kHz.

Although comparatively low entropy was found for many combinations, five MFCCs at 4 kHz bandwidth was the optimum. It is inevitably the case that using more than nine MFCC coefficients is of a disadvantage. This means that the typically applied 13 MFCCs in speech and speaker recognition lead to a decreased vowel recognition performance (detailed test in experiment II). Figure 1 (top) also reveals that at a 4 kHz bandwidth, entropy for MFCC 1 was lower than for MFCC 13 which was close to chance level. This means that the energy coefficient alone in our set-up contains more information about vowel category than the first 13 MFCC coefficients together. It seems conceivable that the bandwidth at which recognition was best is around 4 kHz as this is the band that contains most of the vocalic information. It remains unclear, however, whether this finding can be generalized across vowels with high and low f_o ranges. This was tested in experiment II.

3. Experiment II: Impact on f_o variability on recognition performance

We studied the impact of increasing f_o variability in the classification data by creating different data subsets. These subsets were created (a) by adding vowels with increasingly higher f_o to a low f_o range dataset (198–294 Hz) until we arrive at the full range (198-698 Hz) and (b) by taking away low f_o vowels from the full set, until we arrive at a set with a high f_o range (587-698 Hz). We used novel (experiment I: five MFCCs with 4 kHz bandwidth) and traditional MFCC settings (13 MFCCs with varying bandwidth) for the recognition. Additionally, we studied the influence of vocal effort (low,

Table 1. f_o range, f_o levels (see *Speakers and recording procedure*), number of vowel recordings in range (N), and range labels (increasing upper f_o range, IUR; full range, FR; decreasing lower f_o range, DLR).

Experiment	f_o range (Hz)	f_o levels	N	Label
II	196-293	4	378	IUR1
	196-392	8	659	IUR2
	196-494	10	847	IUR3
	196-587	12	1036	IUR4
	196-698	14	1282	FR
	294-698	10	904	DLR1
	392-698	7	623	DLR2
	494-698	5	435	DLR3
	587-698	3	246	DLR4

medium and high) on vowel classification at the full range of f_o using the best parameter setting found in the experiment I (five MFCCs at 4 kHz).

3.1 Method

Recognition methods (unsupervised k-means clustering) as well as performance measure (entropy) were identical as in experiment I with the exception that the dataset was subdivided into the subsets presented in Table 1.

Recognition was carried out in three different settings:

(a) with the number of coefficients and the signal bandwidth which showed the best mean performance in experiment I (five MFCCs at 4 kHz bandwidth) (Fig. 2: Red box plot);
(b) with the number of MFCCs that is most commonly used in speech and speaker recognition, i.e., 13, and the lowest bandwidth found in previous studies (8 kHz; de Wet *et al.*, 2004) (Fig. 2: Green box plot);
(c) like (b) but with the bandwidth that performed best in experiment I (20 kHz). Since many studies do not specify signal bandwidth, we applied the best bandwidth setting from experiment 1 at 13 MFCCs (Fig. 2: Blue box-plot).

3.2 Results and discussion

Figure 2 contains the distributions of results for 500 experimental repetitions for all f_o subsets tested with the best setting from experiment I (five MFCCs at 4 kHz signal bandwidth) and two previously used settings (13 MFCCs and either 8 or 20 kHz bandwidth). The grey area in the background marks the ranges of possible results that could have been obtained with 13 MFCCs, when signal bandwidth was varying between 8 and 20 kHz. It is apparent from the graph that entropy was lower in any of the f_o subsets using the parameters obtained from experiment I, compared to previously used settings

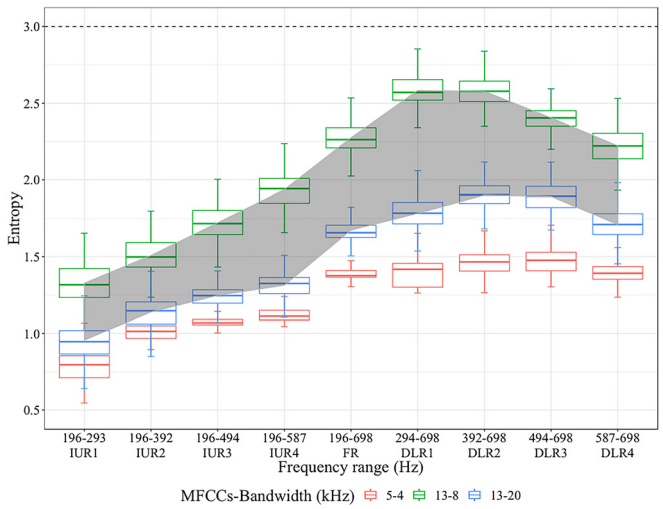

Fig. 2. (Color online) Distributions of entropy (y axis) from 500 experimental repetitions for three MFCCs-bandwidth settings for nine frequency ranges (x axis). Upper and lower ends of vertical bar = range, box = interquartile range, horizontal line = median. The grey shade indicates the performance difference between 8 and 20 kHz bandwidth in 13 MFCCs.

with 8 or 20 kHz bandwidth. It is also apparent that 13 MFCCs applied to a 20 kHz bandwidth show a much higher performance compared to an 8 kHz bandwidth. All in all, this means that unsupervised vowel classification can be drastically enhanced when the right combination between the number of MFCCs and signal bandwidth is chosen for a particular classifier. Most importantly, this combination also works best with a more conventional f_o range (198–293 Hz), which means that it is the best choice for typical speech recorded under laboratory conditions.

Comparing the frequency ranges, we found that the inclusion of higher f_o vowels (IUR1 to FR) decreased recognition performance, in particular when the last two f_o production levels (559 and 698 Hz) were added (FR). This is in line with findings showing that vowel intelligibility decreases with $f_o > 500$ Hz (Sundberg, 2012). However, performance still remains well above chance when best settings (five MFCCs and 4 kHz bandwidth) are chosen. This means that such MFCC representations with the best setting contain sufficient information about the vowel category to allow classification. With previously made choices (13 MFCCs and 8 kHz), performance drops drastically towards chance performance. When lower f_o vowels were removed (FR to DLR4), performance remained rather stable, which means that vowels in an f_o range between 198 and 293 Hz contribute strongest to vowel classification results. This is plausible as this is the range where female f_o is typically most frequent.

The f_o variation in the low and high MFCC ranges (see Fig. 2: IUR1 and DLR4) influences vowel classification performance. Independent of parameter settings, an increase in f_o had a negative impact on classification performance. In contrast to the claims of Ferreira (2007) that "the MFCCs discard pitch information hence the cues are not distracted by pitch harmonics" (p. 2402) we showed that with increasing f_o in the dataset, performance drops. It seems conceivable that the increasing sparsity of the overall spectral shape that goes along with an increase in f_o contains less vocal tract information, and this effect starts from $\sim f_o > 300$ Hz. Hence, undersampling is the most likely reason for the decline in performance starting at $f_o > 293$ Hz. In de Wet *et al.* (2004), the f_o range was between 121 and 246 Hz (including men, women, and children), thus undersampling most probably did not occur. de Wet *et al.* (2004) applied three MFCCs (without energy coefficient) and 13 MFCCs (with energy coefficient) with 8 kHz signal bandwidth. The MFCCs 13 performance was better than MFCCs 3 which was better than chance. Given the current study, performance should increase drastically with a choice of five MFCCs at 4 kHz signal bandwidth. Ferreira (2007) studied vowels with f_o between 98 and 400 Hz (including men, women, and children), using 16 MFCCs (without energy coefficient) and 16 kHz signal bandwidth. Again, performance in this study is expected to improve by using the novel parameter settings obtained here.

We also tested the influence of vocal effort on classification performance at the full range of f_o (500 experimental repetitions for each vocal effort). Results showed a subtle increase with strength of vocal effort (mean entropy at high vocal effort = 1.17, medium = 1.209 and low = 1.29; $F[2,1497] = 309.8$, $p < 0.001$). This is plausible given that articulatory tension is more variable with lower vocal efforts leading to stronger within-category variability. It is also possible that MFCC 1 (the energy coefficient), in particular, contributed to this effect, as energy differences between vowel categories should be most salient at high vocal effort. Most importantly, the magnitude of the effect was low and entropy at all vocal efforts is high above chance level. This means that the effect of vocal effort on undersampled vowel classification can be viewed as secondary in our data.

4. General discussion and conclusion

Humans can categorize vowels up to very high f_o (Smith and Scott, 1980; Maurer *et al.*, 2014; Friedrichs *et al.*, 2015; Friedrichs *et al.*, 2017). Here, for the first time, we showed that machines can perform a similar task with a choice of particular acoustic representation settings for a certain classifier. The results of this study confirm that automatic classification of the wide range of f_o vowels (196-698 Hz) is possible using MFCCs (see Fig. 1). This shows that acoustic information to vocalic category can be reduced to a single MFCC representation even when f_o is high and the vocal tract transfer function is undersampled. Using specific parameter settings (five MFCCs with 4 kHz signal bandwidth), we demonstrated that a drastic performance increase can be gained with a k-means classifier compared to previously used settings (13 MFCCs and 8 kHz bandwidth; de Wet *et al.*, 2004). Other combinations of number of MFCCs and signal bandwidth are applicable but a choice of more than 9 MFCCs (16 MFCCs and 16 kHz bandwidth; Ferreira, 2007) decreased performance at any bandwidth. It is thus apparent that for vowel recognition, the generally applied setting of 13 MFCCs is not recommendable when using k-means clustering. The results suggest that the choice of the number of coefficients in MFCCs seems to

have been underestimated previously. In particular for de Wet *et al.* (2004) and Ferreira (2007) it means that performance could be drastically increased by varying the number of coefficients.

What does this mean for automatic speech recognition? It seems plausible that the high performance at a relatively low signal bandwidth is the result of vocalic information being dominantly present in 4 kHz frequency band. It would be interesting to see whether the results can be generalized for other classification methods as well. Given the overall drastic performance differences in the present study for the five MFCCs at 4 kHz bandwidth using unsupervised k-means classifying, we think that it is plausible to believe that other classifiers might also profit from these MFCC settings. It would further be interesting to test such settings for consonants, in which relevant acoustic cues typically lie much above 4 kHz (e.g., fricatives or plosives). Future research will show whether alternative settings like the ones obtained for vowels in the present study can be of advantage in other recognition situations, especially to improve the accuracy of automatic speech recognition systems in a resource-constrained environment.

Acknowledgments

This work was supported by the Swiss National Science Foundation (SNSF), Grant No. 100016_143943/1. Sandra Schwab contributed in the statistical analysis and Elisa Pellegrino commented on the draft of this paper.

References and links

Brookes, M. (**2011**). "Voicebox: Speech processing toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (Last viewed February 10, 2019).

Choi, W. Y., Song, H. J., Chung, H., Kang, J., and Park, J. G. (**2016**). "I-vector based utterance verification for large-vocabulary speech recognition system," IEEE, ICCCI 2016, pp. 316–319.

de Wet, F., Weber, K., Boves, L., Cranen, B., Bengio, S., and Bourlard, H. (**2004**). "Evaluation of formant-like features on an automatic vowel classification task," J. Acoust. Soc. Am. **116**, 1781–1791.

Ferreira, A. J. S. (**2007**). "Static features in real-time recognition of isolated vowels at high pitch," J. Acoust. Soc. Am. **122**(4), 2389–2404.

Friedrichs, D., Maurer, D., and Dellwo, V. (**2015**). "The phonological function of vowels is maintained at fundamental frequencies up to 880 Hz," J. Acoust. Soc. Am. **138**(1), EL36–EL42.

Friedrichs, D., Maurer, D., Rosen, S., and Dellwo, V. (**2017**). "Vowel recognition at fundamental frequencies up to 1 kHz reveals point vowels as acoustic landmarks," J. Acoust. Soc. Am. **142**(2), 1025–1033.

Goldstein, U. (**1980**). "An articulatory model for the vocal tracts of growing children," D.Sc. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Guevara-Rukoz, A., Mazuka, R., Thiollière, R., Martin, A., Schatz, T., Cristia, A., and Dupoux, E. (**2017**). "Are words in infant directed speech easier to learn? a corpus study of acoustic clarity and phonological density," arXiv:1712.08793.

Hess, W. (**1983**). *Pitch Determination of Speech Signals—Algorithms and Devices by Wolfgang Hess* (Springer, Berlin).

Joliveau, E., Smith, J., and Wolfe, J. (**2004**). "Vocal tract resonances in singing: The soprano voice," J. Acoust. Soc. Am. **116**, 2434–2439.

Kiefte, M., Nearey, T. M., and Assmann, P. F. (**2013**). "Vowel perception in normal speakers," in *Handbook of Vowels and Vowel Disorders*, edited by M. J. Ball and F. E. Gibbon (Psychology Press, New York), pp. 160–185.

Maurer, D. (**2016**). *Acoustics of the Vowel—Preliminaries* (Peter Lang, Bern), pp. 170–182.

Maurer, D., Heureuse, C., Suter, H., Dellwo, V., Friedrichs, D., and Kathiresan, T. (**2018**). "The Zurich corpus of vowel and voice quality, version 1. 0," in *Proceedings of Interspeech 2018*, Hyderabad, India (September 2–6), pp. 1417–1421.

Maurer, D., and Landis, T. (**1995**). "F0-dependence, number alteration, and non-systematic behaviour of the formants in German vowels," Int. J. Neurosci. **83**, 25–44.

Maurer, D., Mok, P., Friedrichs, D., and Dellwo, V. (**2014**). "Intelligibility of high-pitched vowel sounds in the singing and speaking of a female Cantonese Opera singer," in *Proceedings of Interspeech 2014*, Singapore (September 14–18), pp. 2132–2133.

Smith, L. A., and Scott, B. L. (**1980**). "Increasing the intelligibility of sung vowels," J. Acoust. Soc. Am. **67**, 1795–1797.

Sundberg, J. (**2012**). "Perception of singing," in *The Psychology of Music* (Academic, London), Vol. 3, pp. 69–106.

Tan, P. N., Steinbach, M., and Kumar, V. (**2005**). "Cluster analysis: Basic concepts and algorithms," in *Introduction to Data Mining* (Pearson, London), Vol. 1, No. 8, pp. 487–568.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., and Amano, S. (**2007**). "Unsupervised learning of vowel categories from infant-directed speech," Proc. Natl. Acad. Sci. **104**(33), 13273–13278.

Weinstein, E., Mengibar, P. J., and Johan, S. (**2016**). "Context-based speech recognition," U.S. patent Specification No. US9311915B2.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (**2006**). "The HTK Book (HTK Version 3.4.1)," Engineering Department, Cambridge University.

Zheng, F., Zhang, G., and Song, Z. (**2001**). "Comparison of different implementations of MFCC," J. Computer Sci. Technol. **16**(6), 582–589.