Evaluation of an automated formant estimation procedure with optimized formant ceiling

Anna Ericsson

Department of linguistics
Magister Thesis 15 HE credits
Phonetics
Spring term 2020
Supervisor: Włodarczak Marcin

Stockholms universitet

Utvärdering av en automatisk formantmätningsprocedur med optimerat formanttak

Anna Ericsson

Sammanfattning

Denna studie utvärderar en automatisk formantmätningsprocedur utvecklad för anpassning efter talare och variationer i tal. Anpassningen åstadkoms genom att använda det formanttak som uppvisar minst variation (i mätningar av F_1 och F_2 i kombination) som det optimerade taket. Denna optimering ger bästa möjliga estimeringar utifrån data, därför skulle troligtvis anpassningen även kunna ske till variation såsom hög f_o. Proceduren har inte utvärderats genom att använda material med kända formantfrekvenser, varför det görs här. Formantmätningsprocedurens prestation testas genom jämförelse med gängse procedur med fasta formanttak, baserade på skillnader mellan kön. Formantmätningarna utförs på syntetiska vokalexemplar, systematiskt varierade i formantfrekvenser och i f_o för att motsvara naturlig variation inom vokaler och mellan talare. Formantmätningarna jämförs mot ursprungsvärdena, procedurerna sinsemellan och med tidigare studier. Resultatet visar att formantmätningsproceduren med optimerat formanttak inte presterar bättre än den gängse proceduren. Båda procedurer presterar bättre än tidigare metoder, men ingen hanterar hög f_o på ett tillfredställande sätt.

Nyckelord

Formantmätningsmetod, Formanttaksoptimering, Hög grundton, Utvärdering

Abstract

This study evaluates an automated formant estimation procedure designed to adapt to speakers and variations in speech. The adaption is achieved by using the formant ceiling with the least variation (in combined estimates of F_1 and F_2) as the optimal ceiling. This optimization renders the best possible estimations given the data, therefore it could presumably also adapt to variations such as high f_o. The procedure has not been evaluated by using material with known formant frequencies. Therefore, this is done here. The performance of the procedure is tested through comparison with a common procedure with fixed ceilings, based on speaker sex. The estimations are carried out on synthetic vowel tokens, systematically varied in formant frequencies and in f_o, to match the natural variation within vowels and between speakers. The formant estimations are compared to target values, compared between procedures and to earlier studies. The results reveal that the formant estimation procedure with optimized ceilings does not perform better than the common procedure. Both procedures perform better than earlier methods, but neither deals satisfactorily with high f_o.

Keywords

Evaluation, Formant estimation, Formant ceiling optimization, High fundamental frequency

Contents

1 Background

Throughout the history of speech studies, a lot of interest has been directed towards the properties of speech sounds, how they are articulated, how they differ from each other and how this can be captured, measured, illustrated, reproduced and manipulated, all with the purpose of trying to better understand how speech works. In the field of speech research, it is established that vowel sounds are deciphered and perceived by identifying the resonance frequencies of the vocal tract known as formants, or by the general formant contour of the speech signal (Monsen & Engebretson, 1983). Formants originate from the shape and form of the vocal tract and can be measured and compared. But vocal tracts vary within and between speech sounds produced by one speaker and also between speakers, making measurements and comparisons a tricky business. This means that measurement techniques need to be reliable in finding formants in highly varied material. As speech and speakers vary so much, a method is needed that is not suited just to a particular kind of speaker, a method that is just as good at measuring whatever kind of speech or speaker variation that might occur. Also, it is necessary today that the method is automated and applicable to large data sets. The study of speech is often more straightforward on male voices. When investigating female or children's voices, which have a considerably higher fundamental frequency, the problem with sparser harmonics in the signal arises, making formant estimations much harder. This is due to the fact that the harmonics are multiples of the fundamental frequency, making lower fundamental frequencies denser in harmonics than higher fundamental frequencies, and therefore also "richer" in information in comparison. Different methods have been developed to deal with this problem with various results (Atal & Hanauer, 1971; Högberg, 1997; Traunmüller & Eriksson, 1997; Acero, 1999; Xia & Epsy-Wilson, 2000; Watanabe, 2001; Dissen, 2019; Granqvist, 2020). One promising method is the optimized formant ceiling method (Escudero et al., 2009). Because of its design to adapt to the speaker by the taking into account variation between individual speakers and specific vowels, this method should better deal with such variation. Assuming that estimation errors in tokens with high fundamental frequency are due to both sparse harmonics, and to some extent mismatch between generic estimation settings and actual vocal tract configuration, the optimized formant estimation procedure should reduce errors even at high fundamental frequencies. The formant ceiling is the highest frequency of the highest measured formant and a number of formants are fitted under this frequency. With common default settings, used in speech analysis programs such as Praat, the ceiling is fixed (or specified in advance as one of two ceilings based on speaker sex). With the optimized formant ceiling the adaption to the speaker is achieved by estimating formants with a range of ceilings and then selecting the ceiling that has the smallest variance (within a single vowel produced by a single speaker) as the optimal ceiling. This method has not fully been evaluated yet, since it has not been tested on material with known formant values. Therefore, the purpose of this thesis is to perform an evaluation of the optimized automated formant estimation procedure introduced by Escudero and colleagues (2009), by comparing the automatic formant ceiling optimization with common default settings based on speaker sex. Testing both formant estimation procedures on synthetic vowel material that is systematically varied, taking into account variation in vocal tract length and configuration resulting both from different speakers and vowels as well as different fundamental frequencies, should reveal the limits and usefulness of the formant ceiling optimization procedure compared to common default settings.

1:1 What are formants?

It is important to note that the term formant can be used in different ways. It can be used to refer to either a property of the vocal tract or to a property of the sound produced by the vocal tract. The definition that is the most common and accepted view (and also the oldest), is that formants may be defined as properties of the vocal tract itself (Stevens & House, 1955; Fant, 1960; Pickett, 1980). According to the other definition, the term formant may refer to a property of the acoustic signal. In this view formants are defined as spectral peaks, that is, places along the frequency scale where harmonics are enhanced due to the resonance properties of the vocal tract configuration. Whichever way one chooses to describe formants, it is generally confirmed in the field of speech research that "vowel sounds can be perceived and decoded by locating the frequencies of the formants or by reference to the overall formant pattern" (Monsen & Engebretson, 1983, p. 89).

To provide an explanation of what formants are, some acoustic background will follow starting by exemplifying the vocal tract as a tube open at one end and closed at the other end. Resonances occur in a tube, which means there will be places of maximum and minimum movement/velocity in the oscillations of the air particles in the tube. These resonances are present as kinetic energy in standing waves at each resonance frequency and each standing wave has places of maximum (anti-node) and minimum (node) of movement/velocity of the particles. The air particles oscillate, moving towards and from each other, making the wave propagate in a motion driven by the variations in high and low pressure between the particles. Therefore, where there is minimum of movement/velocity there is also maximum of pressure (node). Each resonance corresponds to a standing wave between the voice source (at the vocal folds in the glottis) and the mouth opening (at the lips) in the vocal tract, or in a tube between the closed end and the open end of the tube. How resonances in a tube are affected by changes in the form, of the tube is described by the perturbation theory (Chiba & Kajiyama, 1941). If the tube is widened or tightened in some region, the resonance frequencies will rise or fall, depending on where along the tube the change in diameter happens. If a constriction is made in a place of the tube where there is maximum velocity (anti-node) this will cause the resonance frequency to fall and if the constriction is in a place of maximum pressure (node) it will cause the resonance frequency to rise. These resonance frequencies are the cause of the spectral peaks, that is, the formants. For example, both F_1 and F_2 are low in the vowel /u/ because of the constriction at the lips where both F_1 and F_2 have a maximum in velocity, causing the formants to fall.

Note that formants will be affected differently, in the sense that they have different frequencies and wavelengths, and different wavelengths entail different areas of maximum and minimum of velocity and pressure along the tube. This means that a constriction at a given position can affect different formants differently. This is exemplified in the vowel /i/ which has a high F_2 but a low F_1, because the constriction in the palatal region corresponds to maximum pressure in F_2 (making F_2 high) but the same region corresponds to an area of maximum velocity in F_1 (making F_1 low). In the vowel /ɒ/ the F_2 is also low because of the constriction in the pharyngeal region where F_2 has a maximum of velocity, making F_2 low, but the F_1 in the same vowel is closer to maximum pressure, making F_1 high.

Much of what is established today in acoustic theory of speech can be derived from Fant's work in the 60s (Fant, 1970). The formant pattern is, according to Fant, the set of resonance frequencies of the vocal tract. It conditions the essence of a vowel spectrum and serves as a good correlate to articulatory positions (Fant, 1970). This is often explained through the source-filter model, developed mainly on the basis of the early work by Fant, with substantial contribution by Stevens (Stevens, 2000). The source-filter model presents the sound produced by the vocal chords in the larynx as the voice-source. The vocal tract is described as a filter that can change its shape and thereby its filter properties. The resonance properties of the vocal tract will modify the spectral properties of the voice source in such a way that some harmonics will be amplified and some will be dampened. These resonances can be seen as amplitude peaks in the spectral envelope of the signal in frequencies that are enhanced. They can also be visualized as darker areas in a spectrogram, which is a spectral representation of a signal over time. These peaks in a spectrum, or darker areas in a spectrogram are graphic representations of the resonances that are called formants in speech. Note that resonances will occur in any tube, even if it is a straight, uniform tube. The resonances that such a uniform tube causes correspond to a neutral vocal tract which in turn corresponds to the pronunciation of the vowel schwa (see Figure 1a and b). The vowel schwa is a neutral vowel sound, meaning that articulators are in their resting positions.

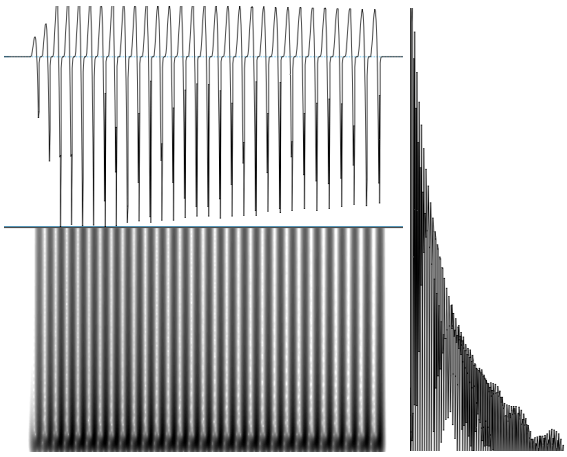


Figure 1a. A synthesized voice source signal (i.e. a signal without resonance peaks). Sound wave (top) and spectrogram (bottom) with corresponding spectrum (right).

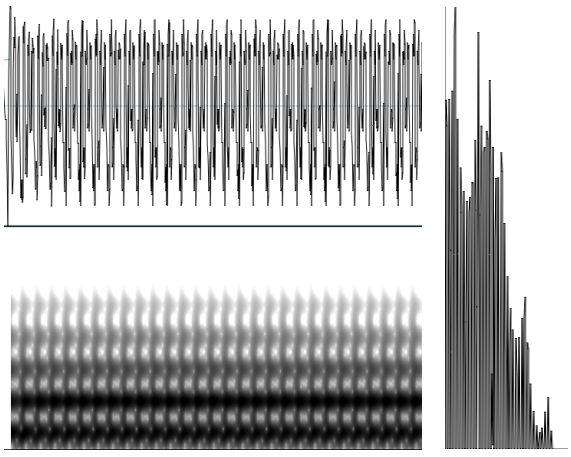


Figure 1b. A synthesized signal with resonances caused by a neutral tube (the resonances of the neutral vowel schwa). Sound wave (top) and spectrogram (bottom) with corresponding spectrum (right).

As mentioned above, making different constrictions in a tube will affect the resonances of the tube. This is equivalent to changing the form of the vocal tract by articulation into different shapes, which results in different resonances. It is these resonances that correspond to the formants of specific vowels, which give the vowels their characters. The precise frequencies at which these formants occur depend on the form and also the size of the resonator, therefore, if the resonator is a tube, the form and length of the tube affects the resonances. In basic terms this can be described as for example when articulating the vowel /i/, the tongue's position is forward and upward towards the upper front teeth. This means that the vocal tract is formed into a tight tube (constricted) in the front of the mouth and a wide tube (widened) towards the back of the vocal tract. This form gives the resonances a low first formant and high second formant. When the form of the vocal tract is changed this way, this will change the resonances, as seen in Figure 2.

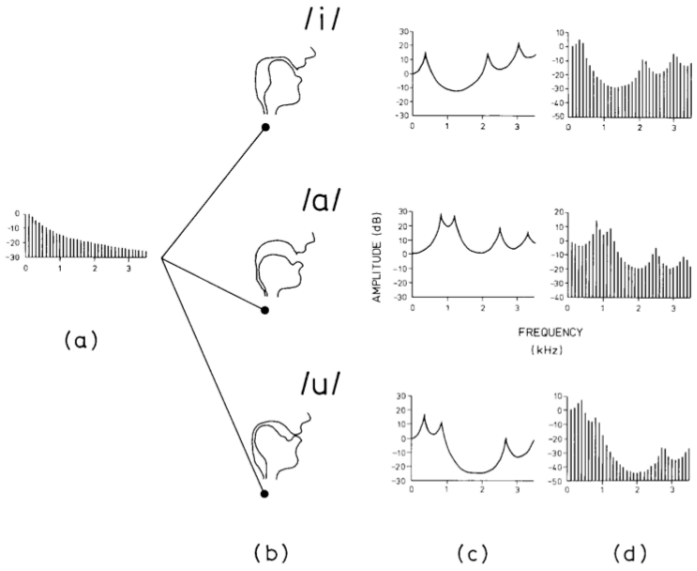


Figure 2. Illustration from left to right of (a) a spectrum of a voice source (not yet affected by the resonance properties of a tube), (b, c) the form and resonance properties of the vocal tract (or a tube) for the three vowels /i/ /ɑ/ /u/ and (d) spectrums of the output vowel sounds as they would look modified by the tube or vocal tract forms for the different vowels. Figure from P. J. Bailey, (1983), with permission from the author.

The resonance properties of different tubes can also be described by the source-filter theory. The output speech signal is the result of a sound source and the vocal tract seen as a filter, and the resonances are determined by the properties of the filter (Fant, 1960). This way the source-filter theory serves as a good model for showing how vowels correspond to the different forms of a tube, based on their articulation. Since the specific form of the vocal tract (determined by how the tongue, jaw and lips are positioned) determines the frequencies at which the formants occur, the peaks roughly describe and characterize a vowel in terms of the peaks in the spectral envelope of the acoustic signal. Formants correspond to the articulation in the way that the first formant (F_1)[1], the lowest resonance, rises in frequency as the speaker lowers the jaw, which includes a lower tongue position. The second formant (F_2) rises as the speaker moves the tongue

[1] *The formants are named after how they appear along the frequency scale so that the first formant (F_1) is the lowest resonance and the second formant (F_2) is the next resonance and so on (Titze et al., 2015).*

body forward towards the teeth. The connection between F_1 and the vertical position of the jaw as well as the connection between F_2 and the horizontal position of the tongue, can be illustrated by the vowel chart of the International Phonetic Association's alphabet (IPA, 1999; see Figure 3). The vowels are shown with their relative F_1 and F_2 values in the vowel chart, corresponding to the vowel space. The vowel chart can be seen as a visualization of the oral cavity, where the position for each vowel in the diagram refers to the tongues vertical and horizontal position, like coordinates in the oral cavity during pronunciation of that vowel. The illustration shows the two first formants, which are most important for carrying the information about the vowel identity, giving each vowel its specific character. In higher formants the correspondence to articulation is not as straight forward, but the third formant (F_3) is related to the position of the tip of the tongue and the small cavity under the tongue (Sundberg 2001) and even higher formants carry mostly other types of information such as for voice quality (Sundberg 2001).

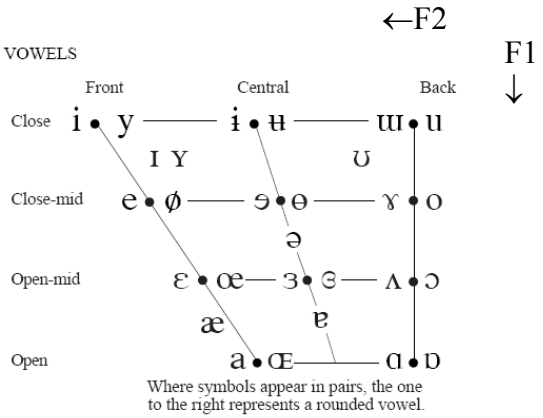


Figure 3. Vowel chart from IPA (International Phonetic Association), with attached arrows for formant 1 and formant 2, illustrating how relative formants values corresponds to specific vowels.

Formants differ not only between vowels but they also differ between speakers. For example, it has long been known for that the same vowel produced by different people will have different formant frequencies (Peterson and Barney, 1952). Differences between speakers are caused by differences in individual articulation (because of context or language etc.) but also by differences in speakers' vocal tract size. As speakers naturally come in different shapes and sizes, so do their vocal tracts. Children have smaller vocal tracts than adults, and women mostly have smaller vocal tracts than men. This size difference affects the speech signal of vowels, in that the smaller the resonating vocal tract is, the higher formants or resonances can be expected, just as a small piccolo flute resonates with higher frequencies than a large flute does. This is because in a larger/longer tube (vocal tract), the distance between nodes and anti-nodes will be longer and therefore resonance frequencies (formants) will be lower.

It is important to bear in mind that this work is in the field of phonetics and that the description of formants here, is highly influenced by the way the term is used this field. The term "formant" can also be used when talking about resonances in a room and it can be used sometimes synonymously with "resonance" and also to some extent with "pole". As mentioned earlier, some researchers use the term formant to mean a peak in the spectral envelope (as a property of the sound of the voice), others to mean the resonance of the vocal tract (as a property of the vocal tract) and to others it means the pole in a mathematical filter model (a property of a mathematical model).

1:2 Estimating formants

Estimating, tracking or measuring formants has been given a lot of attention in speech analysis and speech recognition. Since formants have such an importance for determining the phonetic content and its close connection to the vocal tract, formant frequencies are a desirable and legitimate measure to use. But reliable formant frequencies are difficult to extract from the speech wave and therefore many methods have been developed to try to do this, with various results.

The traditional way to find formants has been to visually locate, with the human eye, the broad peaks caused by the form of the vocal tract in the speech signal by looking at broadband spectrogram or spectrum. The spectrogram is a three-dimensional representation of frequency distribution over time, with time on the x-axis, frequency on the y-axis and amount of energy or amplitude in color saturation. The spectrum, showing frequency on the x-axis and amplitude on the y-axis, can show only a precise moment in time (or a mean over a period of time). A spectrogram can therefore be seen as several spectra over time. The darker areas in the spectrogram and the broad peaks in the spectrum reveal the resonances.

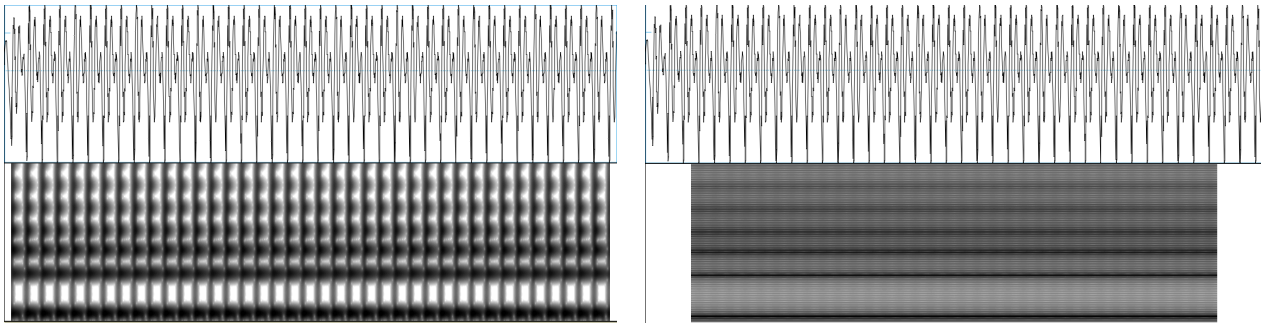


Figure 4a. Spectrograms and waveforms of a synthetic vowel /i/ showing broad band spectrogram to the left and narrow band spectrogram to the right with waveforms above. Time (x-axis), frequency (y-axis) and amplitude (color saturation) in spectrograms. Time (x-axis) and amplitude (y-axis) in waveforms.

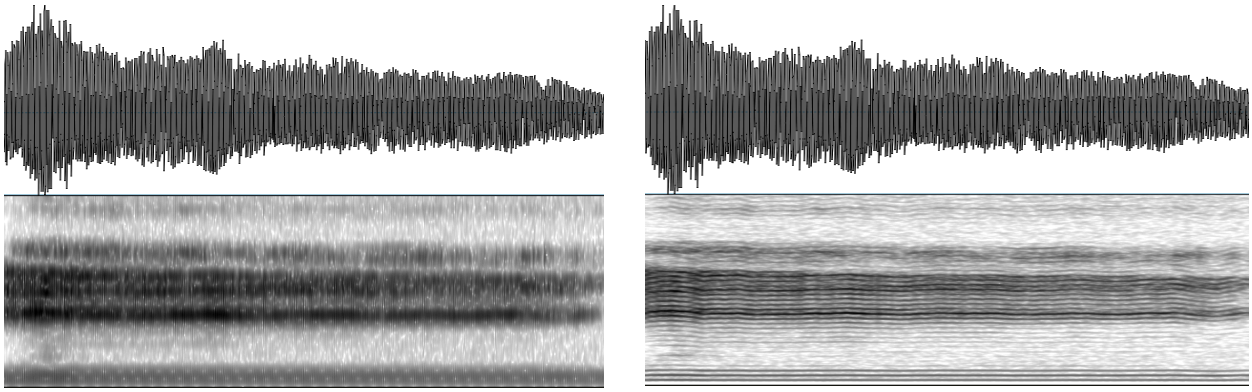


Figure 4b. Spectrograms and waveforms of a natural vowel /i/ showing broad band spectrogram to the left and narrow band spectrogram to the right with waveforms above. Time (x-axis), frequency (y-axis) and amplitude (color saturation) in spectrograms. Time (x-axis) and amplitude (y-axis) in waveforms.

Both the spectrogram and the spectrum need some smoothing to avoid confusing these broad peaks with single partials or harmonics, which appear as narrower peaks. The smoothing of both is needed but can result in some problems, for example, it may cause peaks that lie close to each other to be hard to separate, in both spectrum and spectrogram. The smoothing is done by looking at broadband (as opposed to narrowband) spectrogram or spectra. A narrow band spectrogram will show separate harmonics (see Figure 4a and b). In a spectrum or spectrogram, smoothing can be achieved by using a shorter spectral analysis window. There is a tradeoff between time and frequency resolution in the sense that if is used a shorter analysis window, the time resolution will be very precise/fine scaled and the frequency more blurred. Therefore, in the broad band spectrogram (i.e., using a short analysis window), the fine time resolution and the smoothing in frequency needed when finding resonances, is precisely the result you get. When locating formants, spectrum and spectrogram can be used simultaneously (as well as narrow and broadband spectrograms). But manually reading a spectrum or spectrogram relies on an experienced reader with phonetic expertise and is a time-consuming method. Nowadays this method is not used very often, since it can be done easier, more quickly, and often more accurately, in automated ways (Monsen & Engebretson, 1983; Wood, 1989).

The method that has been most used for automatic formant estimation is linear predictive coding (LPC). Estimating formants with LPC can be done with several speech processing programs such as Wavesurfer (Sjölander & Beskow, 2000) and Praat (Boersma, 2002). LPC is done by means of a predictive coding algorithm. With the mathematical operation of linear prediction, future values of a discrete time signal are estimated as a linear function of previous samples. This will define a filter (find the broad spectral peaks in the signal) which approximates the vocal tract filter function (i.e., the vocal tract resonance properties). In other words, LPC estimates the underlying resonance frequency, that is revealed by the temporal response from the vocal tract, providing an accurate representation of the vocal tract filter function. It is a method that works well for non-nasalized vowels, as it assumes that the voice spectrum is shaped by broad spectral peaks with no prominent valleys (anti-formants). But for nasals, laterals and fricatives, in which valleys are important, it does not work as well. When using LPC the number of peaks to find in the spectrum has to be specified in advance, since the number of peaks that LPC will fit to the spectrum is specified by the number of coefficients in the linear predictive equation (if too few are specified, LPC will fail to register peaks that actually exist in the spectrum). In practice the number of formants expected depends on the frequency range and the length of the speaker's vocal tract. Therefore, the predictions used in the calculations of the vocal tract filter function with LPC make the method somewhat suboptimal when it is used for speakers with unexpected vocal tract length, as formants in such cases might not fit into the preselected frequency range. It is also worth noting that there are several different applications of LPC within the area of estimating formants. LPC is used, for example, in formant estimations with dynamic programming (Xia & Epsy-Wilson, 2000), in formant estimations using combination of filter bank and cepstral coefficients (Högberg, 1997) and in estimating formants with hidden Markov models (Acero, 1999).

There are other types of formant estimation methods based on inversed filtering. Mostly, inverse filtering has been used to subtract vocal tract resonances to get to the glottal wave of the voice source, but it can also be used to get to the vocal tract resonances and formants. This has been done using programs such as the De Cap and Sopran software (Granqvist, 2020). In short, it filters the audio signal with the inverse of the estimated transfer function, eliminating the effect of the vocal tract transfer function on the input signal. If this is done correctly it will result in the clean glottal voice source. Hence, the eliminated spectral peaks (transfer function)

should correspond to the proper formants. But this procedure involves setting filters and bandwidths correctly which is done manually and therefore requires training and expertise.

The inverse-filter control (IFC) is another inverse filtering method (Watanabe, 2001). The method separates the speech signal by controlling inverse filters and estimates formants as mean frequencies of separated waves. The method adapts number of formants (and bandwidths) to be estimated according to the signal. When the number of formants assumed with the method equals to that of relatively clear resonances in the signal it can estimate this fixed number of formants. The method seems promising and it was observed to show fewer errors than LPC, especially when analyzing real speech. But it does not work as well as LPC on synthetic speech, even though errors have been reported to be small (Watanabe, 2001). This method does not require the same degree of expertise and manual workload as the De Cap and Sopran does.

Another method to estimate formants automatically that is a rather recent approach is to use deep learning networks to estimate formants. By using supervised machine learning techniques, networks are trained on for example annotated corpus of read speech. Dissen et al. (2019) used LPC-based cepstral coefficients and raw spectrogram in their input and their networks performs well in comparison with other methods. Dissen and colleagues also proposed and evaluated a change in their network architecture that allows to adapt the models to new domains of speaker types with different speaker characteristics and speaker styles and the adapted networks improved further in estimation and tracking of formants (Dissen et al., 2019).

Also worth noting is a way to check and refine the results of formant estimations is by using analysis by synthesis (A-b-S), which has been used in several studies (Atal & Hanauer, 1971; Traunmüller & Eriksson, 1997). This approach involves synthesizing vowels with the estimated formant frequencies to check the result. Often a perception test is included, evaluating how good or natural the material sounds, to get a measure of how good the formant estimations were in the first place.

1:3 Estimating formants with high fundamental frequency

When measuring formants, it is essential that there is enough information in the signal. In practice this means that the segments need to be voiced. The signal contains more information the lower the fundamental frequency (f_o)[2] is. This is because harmonics in a periodical waveform are multiples of f_o. If f_o is high and the harmonics are therefore much sparser, it makes finding formants in the signal much harder. The accuracy of the formant estimation never can be more precise than the distance between harmonics (Davis & Lindblom, 2000). The lack of information in a signal with higher f_o can give the result that resonance peaks may lie between harmonics or two peaks may lie so close together that one or several peaks may be missed. It can also result in two harmonics being mistaken for two formants. The problem with finding formants in signals with high f_o affects most kinds of formant estimation methods, whether manual or automated.

High f_o can be found in a number of situations, for example, in singing, when speaking with a raised voice, when talking to small children, and also in children's speech. In singing, f_o can be

[2] The abbreviation for fundamental frequency is written f_o, short for first oscillation, (Titze et al., 2015).

as high for male tenor voices as 523 Hz and for female soprano voices 1046 Hz, corresponding to their "high C" (Sundberg & Skoog, 1995). Such high f_o makes it problematic to assess formants. In some cases, f_o is actually higher than F_1, affecting vowel intelligibility (Sundberg, 2001). Under such conditions the strategy of the singer to increase intelligibility, seems to be to raise F_1 to a frequency above the frequency of f_o by lowering the jaw (Sundberg, 1975; Sundberg & Skoog, 1995; Sundberg, 2001; Deme, 2014).

When it comes to speaking with a raised voice, increasing vocal effort from neutral to loud increases mean f_o as well (Jessen et al., 2005). The f_o increases in mean from neutral speech (when speaking and reading with normal voice) to loud speech, (when speaking and reading with Lombard settings 80 dB_{SPL} white noise in headphones) with 33-39 Hz (Jessen et al., 2005). The primary tool for raising sound level in phonation is to increase subglottal pressure according to Ladefoged (1962), and in untrained voices, f_o tends to follow voice effort as a natural physiological/acoustic consequence of increased subglottal pressure (Gramming et al., 1988; Titze & Sundberg, 1992). Meaning that as a natural consequence of raising the voice (by increasing subglottal pressure) the f_o will also increase, since the vocal folds will open and close in a faster speed at higher air speed/pressure through the glottis. The relation between fundamental frequency and vocal intensity is one octave per 8-9 dB increase (Titze & Sundberg, 1992). When non-singers (with untrained voices) are asked to increase vocal effort they almost always also increase f_o (Sundberg, 2001). In Gramming et al. (1988), mean pitch was found to increase by about a half-semitone (i.e., corresponding to a change in 50 cent, since 100 cent equals one semitone) per decibel sound level (Gramming et al., 1988).

People tend to use high and variable f_o when talking to infants and small infants (Fernald & Simon, 1984; Fernald et al., 1989; Kitamura & Burnham, 2003). Sometimes the mean f_o reaches frequencies as high as over 400 Hz in American English, French and Italian mothers, and close to 400 Hz in British, German and Japanese mothers. French, Italian and British fathers reach mean frequencies over 200 Hz and American, German and Japanese fathers reach over 150 Hz in mean frequency (Fernald et al. 1989). This high (exaggerated) and varied f_o in speech to infants has been suggested to be a primordial quality used by parents to express vocal emotion and to emotionally regulate the infant and keeping its attention. This kind of speech is also believed to have the purpose of sounding non-threatening, communicating non-aggressiveness to the infant (Kalashnikova et al., 2017). This kind of speech comes spontaneously when talking to a small child and is found across language and culture (Fernald et al., 1989; Uther et al., 2007; Broesch & Bryant, 2015), speaker sex (Fernald et al., 1989) and speaker age (Trainor et al., 2000).

Formant estimations are much harder to accomplish for speech signals with high f_o, but they are nevertheless necessary. Therefore, it is important to know how well existing techniques actually handle high f_o when estimating formants. Inverse filtering is one formant estimation method that does handle high f_o, but this procedure requires integrating several signals, such as glottal airflow and the frequencies of the inverse filters and bandwidths are set manually, including manual evaluation and validation that requires phonetic experience (Hertegård & Gauffin, 1993; Sundberg, 2013). Inverse filtering techniques thus involve a heavy manual workload, and automatized procedures to increase formant estimation precision and reliability would be very useful. Analysis by synthesis has also been applied in formant estimations on speech with high f_o (Traunmüller & Eriksson, 1997), in an attempt to reduce errors and testing on synthetic speech. The method seems to work satisfactorily, but only on fully voiced slices of speech that do not include any drastic changes in voicing or formant frequencies. Thus, problems still remain with synchronization (between synthetic versions and the original signal),

with onset and offset of voicing and rapid formant transitions. These problems could be solved by applying the method period by period, but that would require a reliable detection of perfect periods in advance (before applying the method). This seems time consuming and also needs to be tested before saying anything about the methods usefulness in estimating formants in high f_o. Also, this method runs into problems when F_1 and f_o, are close and when f_o is higher than F_1.

One method that has been introduced to improve formant estimation is the optimized formant ceiling procedure, which can be implemented as part of automatized LPC formant estimation (Escudero et al., 2009). It is designed to account for variations in the vocal tract size of the speaker by using the formant ceiling that is best suited for each speaker and sound being estimated. The formant ceiling is the highest frequency at which the highest specified formant is expected to appear and under this frequency a set number of formants will be fitted. With a fixed number of estimated formants, the formant ceiling should be higher for speakers with shorter vocal tracts. In standard formant estimation procedures this is taken into consideration by using generic ceilings based on speaker sex. But these generic ceilings do not account for all the kinds of variation there may be between speakers, nor do they account for variation between different vowels. Between vowels the vocal tract also has different sizes, because of their varying constrictions and because of the shape and size of constrictions and widenings in separate vowels. These variations make the vocal tract configuration differ in size between for example the vowels /i/ and /u/, (Escudero, 2009). With the optimized formant ceiling procedure, formants are estimated with a range of ceilings. The variance in estimations is calculated (summed for F_1 and F_2) and the ceiling that results in the smallest amount of variance is used. This way the optimized ceiling method should be able to account for different vocal tract sizes and configurations. Although it is not specifically designed to tackle the problem of high f_o, the optimized method has been used on speech where f_o is high, as it typically is in speech directed to small children (Wang et al., 2015; Marklund & Gustavsson, *in press*). Because of its adaptation to the speaker it is possible that the method could improve formant estimations also when f_o is high. However, the method has not yet been tested on material with known formant values, such as synthesized speech, which needs to be done to be able to evaluate the methods efficiency with certainty. By making formant estimations on synthesized vowels, with the optimized formant ceiling procedure and with standard formant estimation settings and by comparing the two, this thesis will evaluate the optimized formant ceiling procedure.

1:4 Aims and research questions

The purpose of this study is to evaluate the optimized formant estimation procedure. The procedure has been used on real speech, but never before tested on synthesized material. The procedure has been used on infant-directed speech (where f_o is often high and modulated) and seems to be a promising method to use for this, since it adapts the measurements to capture what is actually there in the signal by testing and using the measurements that have the smallest variance. Comparing the common automatic formant estimation method, which uses fixed settings based on speaker sex, with the automatic formant ceiling optimization procedure, while systematically varying f_o, will hopefully reveal the usefulness and limitations of the procedure. It will hopefully also become clear whether the optimized ceiling procedure from Escudero and colleagues (2009) renders more accurate formant estimations compared with common generic formant ceiling settings (based on speaker sex) and whether these estimations are more accurate when f_o is high. Since the optimized ceiling procedure is designed to adapt more to the signal than the common procedure does, it is anticipated that the formant estimations from the optimized ceiling procedure will prove to be more precise and also to better deal with higher f_o

than the common method. For the same reason, the optimized ceiling procedure will also presumably be better at dealing with different speaker sex and variation within vowels.
In short, this thesis will try to answer the following questions:

- Will the optimized formant ceiling procedure prove to be generally better than the procedure with common generic formant ceiling settings?
  - Will it be better at handling variation between speaker sex?
  - Will it be better at high f_o?
  - Will it be better at handling variations between and within vowels?

2 Method

The method in this study closely follows Monsen and Engebretson (1983), comparing two formant estimation methods using synthesized target vowels. Monsen and Engebretson (1983) compared manual formant tracking with automatic formant estimation (LPC), while the focus in this study is comparing automated LPC formant tracking method with generic formant ceilings to the procedure in which formant ceilings are optimized from Escudero (2009). Synthesized vowels with known formant and f_o values were created in order to allow an evaluation of the accuracy of the two estimation procedures. The set of vowels in the present study includes several tokens per vowel type, varied in F_1 and F_2, simulating the within-speaker variation necessary for the optimal formant ceiling procedure.

2:1 Vowel synthesis

The nine long Swedish vowels i, e, y, ä, ö, u, a, å, o were included, that is: [i, e, y, æ, ø, ʉ, ɒ, o, u]. For each vowel type, 50 F_1-F_2 configurations were specified, 25 based on data from male speakers and 25 based on data from female speakers (see Figure 5). Measures (F_1-F_3), from (Eklund & Traunmüller, 1997) were used, systematically varying F_1 and F_2 according to their respective standard deviations (see Table 1), from -2SD to +2SD in steps of 1SD. F_3 was not varied and the fourth formant (F_4) was set to 4000 Hz for the male formant values and 4500 Hz for the female formant values. The tokens were synthesized for each formant configuration in combination with f_o varying from 100 Hz to 500 Hz in steps of 100 Hz. This was done to evaluate the correctness of the estimations using default formant ceiling settings based on speaker sex (generic formant ceiling) and by using the optimized formant procedure (optimized formant ceiling) and to be able to see how accurately the two measuring procedures would estimate the formants at different f_o. The nine vowel types and their variations gave a total of 2250 different vowel tokens (nine vowel types * two speaker sexes * 25 variations * five f_o levels = 2250 separate tokens).

Table 1. Mean formant frequencies and standard deviations (in hertz) for F_1 and F_2 for male and female vowels from Eklund & Traunmüller, (1997).

Vowel	M F_1 Mean	M F_1 SD	M F_2 Mean	M F_2 SD	F F_1 Mean	F F_1 SD	F F_2 Mean	F F_2 SD
i [i]	291	12	2107	74	351	34	2455	190
e [e]	376	14	2152	41	438	24	2500	178
y [y]	285	4	1988	61	353	26	2319	194
ä [æ]	612	40	1501	79	755	64	1890	171
ö [ø]	436	21	1601	60	517	51	1900	86
u [ʉ]	328	17	1679	52	386	10	1904	84
a [ɒ]	560	41	876	32	665	47	1038	58
å [o]	382	15	642	14	424	20	748	67
o [u]	320	20	639	40	374	16	718	58

Synthesized vowel tokens were created using Praat (versions 6.0.37 and 6.1.09; Boersma and Weenink, 2018). The tokens where created with KlattGrid (pulse source), keeping Praat's standard settings but specifying formant frequencies for F_1-F_4 and varying the f_o. Bandwidths were kept at Praat's standards (i.e. 50 Hz for F_1 and F_2 and 100 Hz for F_3), as was vowel duration (400 ms). To create the vowel tokens and variations, a script was written by Marcin Włodarczak and used in Praat, reading formant values with systematic variations from a csv table.

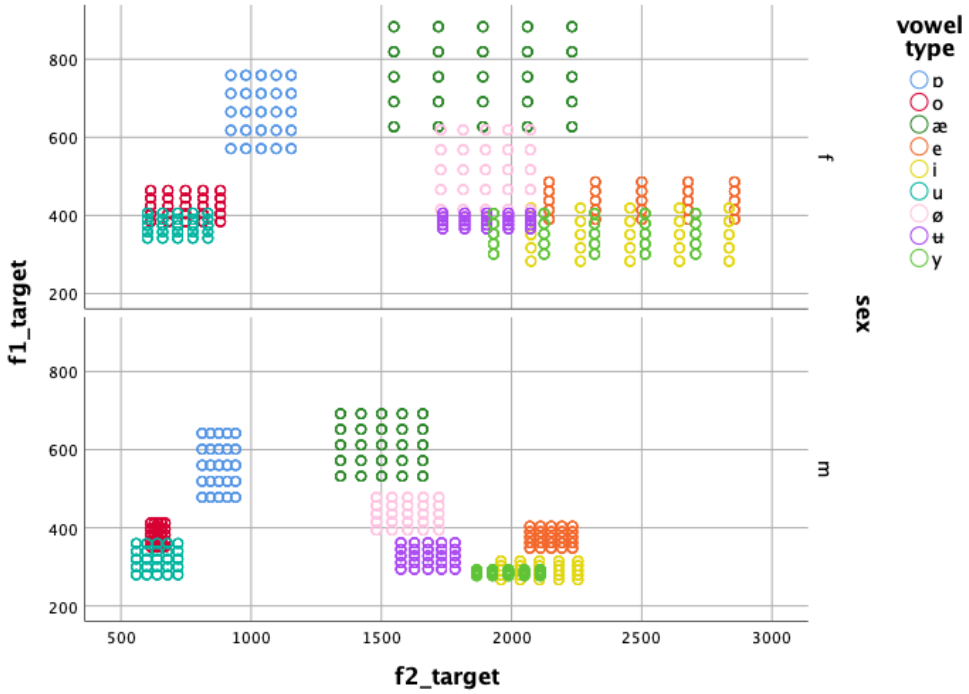


Figure 5. All synthesized vowel types and their variations into tokens, showing F_1 (y-axis) and F_2 (x-axis) in Hertz, split by speaker sex, showing female tokens (top) and male (bottom). The spread is so different between vowel types because their formant frequencies are based on real mean and standard deviation data.

2:2 Formant estimation

All vowel tokens' F_1 and F_2 were estimated using generic defaults and with Escudero's optimized formant ceiling procedure. Except for varying the formant ceilings and increasing number of formants (as it is recommended to adjust the number of formants to the ceiling) all recommended settings were kept. Formant estimations were carried out in Praat (versions 6.0.37 and 6.1.09; Boersma and Weenink, 2018), (Burgh method). The settings used were: Time step (s): 0.0, Maximum number of formants: 6.0, Window length (s): 0.025 and Pre-emphasis from (Hz): 50. The formant ceiling was varied between 4000 Hz and 6500 Hz, in steps of 10 Hz (Escudero et al., 2009). These ceiling settings also cover estimations with the generic ceilings of 5000 Hz for male and 5500 Hz for female sex that can be found from the recommended settings in Praat.

A Praat script was created by Lisa Gustavsson (Anna Ericsson participated in adjusting the script for this study) to estimate formants for every token and ceiling using the settings above. The script calculates four to six formants (F_1-F_6), depending on ceiling. If the ceiling is 4000-4499 Hz, the script will look for 4 formants, if the ceiling is 4500-4999 Hz, it will look for 4.5 formants, when the ceiling is 5000-5499 Hz it will fit 5 formants, and so on, up to maximum ceiling of 6500 Hz, where the script will fit 6 formants. The script then saves values from estimations for each token and ceiling in a text file, rendering 251 text files per token. Out of these formant estimations, only the first two formants were chosen for evaluation in this study.

A script was created by Ellen Marklund (Anna Ericsson participated in adjusting the script for this study) in R 3.5.1 (R Core Team, 2018) to find the optimized ceiling for each vowel token and condition[3] and to collect all estimations of interest in one datafile. The script was designed to read from the text files containing formant estimations. For every combination of token and formant ceiling value, the mean of the formant estimations over the whole duration of the token is calculated. When this is done for all tokens' F_1 and F_2, the script calculates the variance of F_1 and F_2 for each ceiling. Then the script picks the optimal ceiling based on where the logarithm of the variance for F_1 and F_2 combined is the smallest. The script also calculates and picks the mean formant estimation, F_1 and F_2, for the generic ceilings. In addition, the script collects information about vowel type, sex and target formants (i.e. the formant values used as input to the synthesis). The resulting data file contains the following information: name of token (corresponding to conditions sex, vowel and f_0, within the token), speaker sex, vowel type, f_0, F_1 target, F_2 target, F_1 generic, F_2 generic, F_1 optimal, F_2 optimal, generic ceiling and optimal ceiling for every token.

[3] *Condition refers to any combination of vowel type, speaker sex and f_0 (e.g., male /e/ at 200 Hz, or female /i/ at 300 Hz). Each condition thus comprises 25 different tokens, with varying formant values. There is a total of 90 different conditions (2 sexes x 5 f_0 x 9 vowel types).*

3 Results

3:1 Initial analysis

The first step of the analysis was to find the distance between the estimations and targets, to make it possible to see how close the estimations of both methods got to the actual formants. For this measure, the difference between the estimations with optimal ceiling and the target were calculated, as well as the difference between the estimations with generic ceiling and target. Absolute measures were used to see how far from the target estimations were, regardless of whether formants were over- or underestimated. Mean measures with F_1 and F_2 together were calculated to enable a comparison with earlier data from Monsen and Engebretson (1983) (see Figure 6).

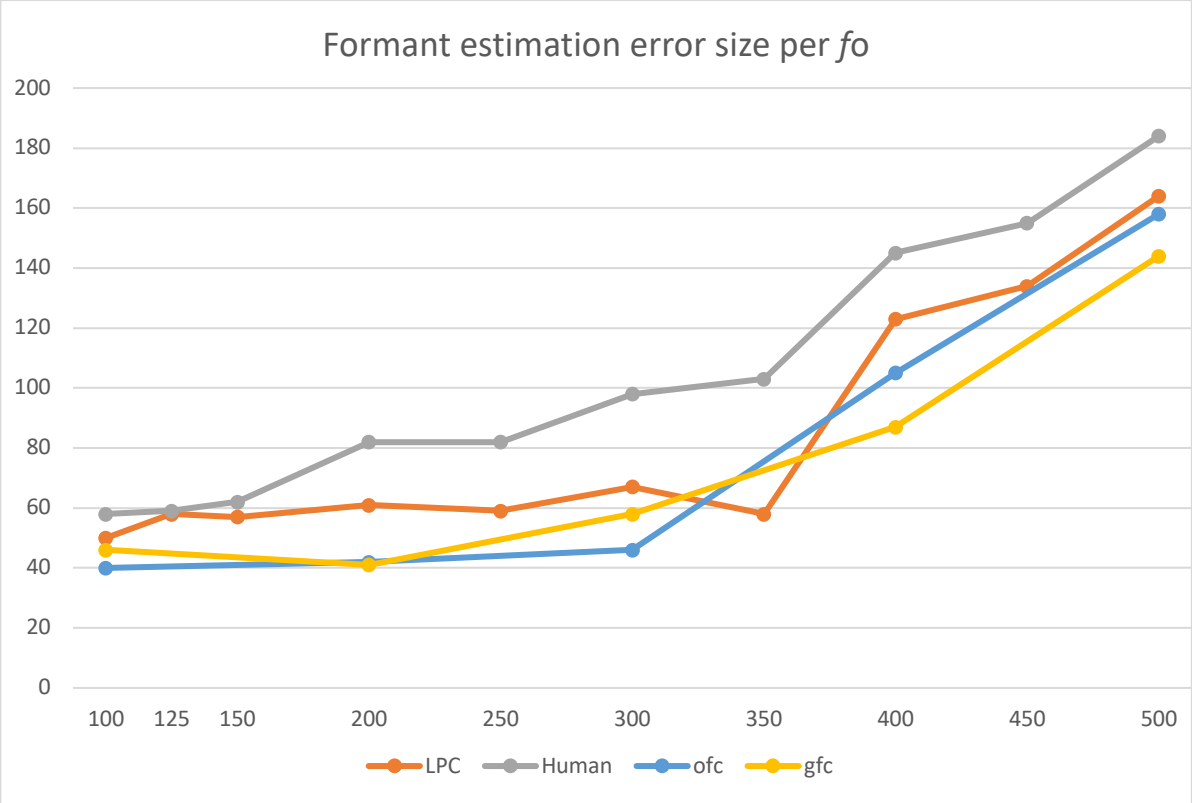


Figure 6. Formant estimations from Monsen and Engebretson (1983) and from current study in absolute mean error from target for all vowel tokens at each f_0 level. Monsen and Engebretson's automated LPC estimations (LPC), Monsen and Engebretson's estimations made by skilled formant readers (Human), current study's estimations with optimal formant ceiling (ofc) and current study's estimations with generic formant ceiling (gfc). Note that f_0 levels are sparser in the current study than in Monsen and Engebretson (1983).

Monsen and Engebretson compared manual estimations with LPC and presented their correctness in terms of mean error from target, including all measurements all together. Therefore, the estimations from the current study are shown here in the same way, with mean

absolute error in estimations from target. This reveals that the best estimations are made by the optimal formant ceiling procedure at low f_o, and with generic ceilings at high f_o. It is clear that both procedures in the current study perform better than Monsen and Engebretson's measures. These results suggest not only that the optimal formant ceiling procedure is a good procedure to use, at least up to 300 Hz, but also that the modern LPC tracking, using generic ceilings is clearly better than the LPC method used by Monsen and Engebretson. The results also reveal that it is still problematic to make formant estimations when f_o is high. In the figure it appears as if one measure at 350 Hz of f_o has surprisingly low error in Monsen and Engebretson's LPC-method. But if this really is better than the two current procedures' estimations at this f_o cannot be answered, since 350 Hz of f_o was not included in the current study.

A log transform of absolute estimation error values was performed, because of an observed skewness in data, to make possible an ANOVA comparison. All statistics were performed in SPSS Statistics 26 (Armonk, NY, USA).

3:2 Statistical analysis

A repeated measures ANOVA was conducted to be able to make comparisons. Within-subjects factors were formant estimation procedure (optimal ceiling, generic ceiling) and formant (F_1, F_2), and between-subjects factors were speaker sex (female and male), vowel ([ɒ] [o] [æ] [e] [i] [u] [ø] [ʉ] [y]), and f_o (100 Hz, 200 Hz, 300 Hz, 400 Hz, 500 Hz). Only main effects and two-way interactions are considered here.

Significant main effects were found for estimation procedure, formant, speaker sex, vowel and f_o. The main effect of formant estimation procedure was significant ($F(1, 2160) = 6.474$, $p = .011$), the formant estimation procedure with a generic ceiling was more accurate overall.

The main effect of speaker sex was significant ($F(1, 2160) = 89.947$, $p < .001$), with male estimations being more accurate than female overall. Interaction between formant estimation procedure and speaker sex was not significant ($F(1, 2160) = .055$), $p = .815$), formant estimation procedures do not perform differently between speaker sex (see Figure 7).

The main effect of formant was significant ($F(1, 2160) = 201.497$, $p < .001$), formant estimations of F_1 are more accurate than estimations of F_2. Interaction effect between formant and speaker sex was significant ($F(1, 2160) = 282.810$, $p < .001$), suggesting that female estimations of F_1 are more accurate than male (see Figure 7, left) but male estimations of F_2 are more accurate than female (see Figure 7, right), although this was not explicitly tested in this study (see the ANOVA table in appendix).

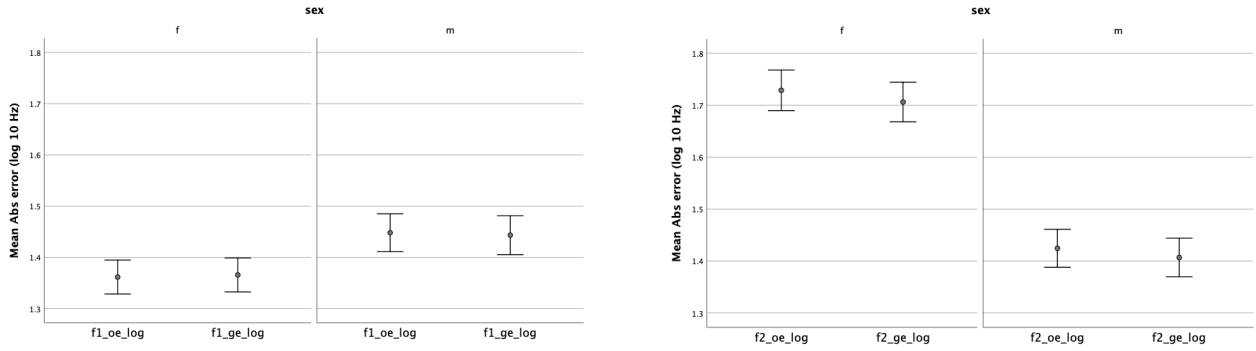


Figure 7. Correctness in estimations of F1 (left) and F2 (right) made with optimal ceiling (f1_oe_log) and generic ceiling (f1_ge_log), for female (f) and male (m) tokens. Bars represent the 95% confidence interval.

Main effect of f_o was was significant ($F(4, 2160) = 941.893$, $p < .001$): estimations become less accurate with rising f_o. Both estimation procedures perform less accurately with rising f_o overall. But a significant interaction effect between formant estimation procedure and f_o ($F(4, 2160) = 10.926$, $p < .001$), indicates that the procedures are different with respect to impact of f_o and as Figure 8 shows, the generic formant ceiling procedure performs slightly better than the optimal formant ceiling procedure with rising f_o, except for at 300 Hz of f_o, where the optimal formant ceiling procedure performs better. Henceforth, for the sake of simplicity, formant estimation procedures will be called optimal procedure and generic procedure but with the knowledge that the ceilings and not the procedures are optimal and generic, respectively.

The interaction effect between formant and f_o was significant ($F(4, 2160) = 35.500$, $p < .001$): both F_1 and F_2 estimations become worse with rising f_o. F_1 is the less affected overall by f_o, except for in the highest f_o of 500 Hz, where F_1 is slightly less accurate than F_2. Correctness of estimations in terms of the amount of error in F_1 and F_2 compared to target at the different levels of f_o, are shown in Figure 8.

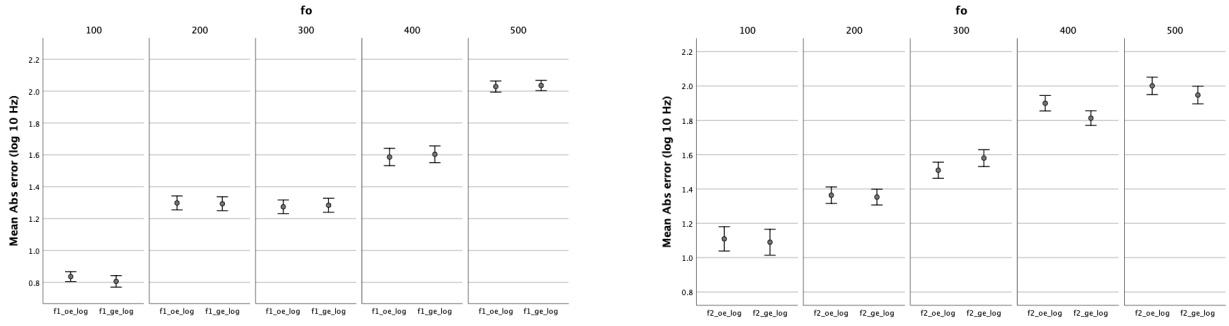


Figure 8. Correctness in estimations of F_1 (left) and F_2 (right) made with optimal ceiling (f1_oe_log) and generic ceiling (f1_ge_log), for each f_o level. Bars represent the 95% confidence interval.

The main effect of vowel was significant ($F(8, 2160) = 38.109$, $p < .001$). Vowel types with the least error (meaning most accurately estimated) are: [ø æ o ʉ ɒ] and the vowels types with the largest error (meaning least accurately estimated) are: [u e y i]. The vowels can be ranked in order from least to greatest error as follows: [ø] [æ] [o] [ʉ] [ɒ] [u] [e] [y] [i]).

The interaction effect between formant estimation procedure and vowel type was not significant ($F(8, 2160) = 1.767$), $p = .079$). The procedures do not perform differently for different vowels. The amount of error in formant estimations varies with vowel type (see Figure 9), but both estimation procedures perform very alike, overall. The interaction effect between formant and vowel type was significant ($F(8, 2160) = 31.181$, $p < .001$). F_1 is remarkably more accurate than F_2 for the vowels [e i u o y]. But as for the vowels [ø ʉ æ ɒ] the difference between F_1 and F_2 is not as clear (see Figure 9).

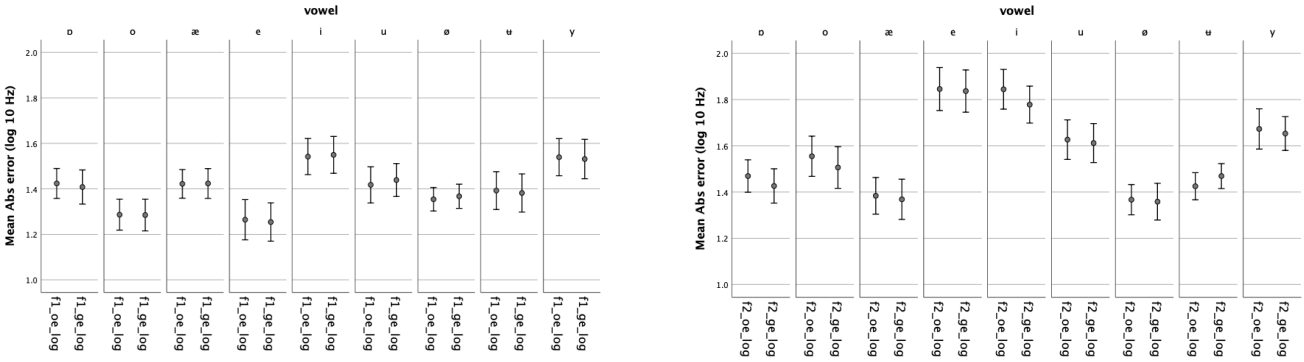


Figure 9. Correctness in estimations of F_1 (left) and F_2 (right) made with optimal ceiling (f1_oe_log) and generic ceiling (f1_ge_log), for each vowel type. Bars represent the 95% confidence interval.

3:3 Exploring differences between estimations and target formants

In Figure 10, estimations of F_1, F_2 and their distance to target values (i.e., the formants specified in the synthetic tokens) at the different f_o levels are shown. This figure, and all subsequent bar diagrams, are made with estimation measures, not log-transformed but expressed in Hz, meaning that they are not linked to the ANOVA but are the underlying measures that the effects shown in the ANOVA are based upon. This is to visualize how far estimations are from the target and in which direction, meaning if they have been over- or underestimated with rising f_o. F_1 estimations are very similar in both procedures and it looks as if F_1 is overestimated with rising f_o. In F_2 estimations some differences can be seen between procedures, but overall, F_2 seems to be underestimated, except for at the highest f_o. Note that the variation is much greater in F_2 than in F_1. An interesting difference occurs at 300 and 400 Hz of f_o, were F_2 is more underestimated with a generic ceiling than with the optimal at 300 Hz, but more underestimated with the optimal ceiling than with the generic ceiling at 400 Hz. At 500 Hz of f_o both procedures instead overestimate the target.

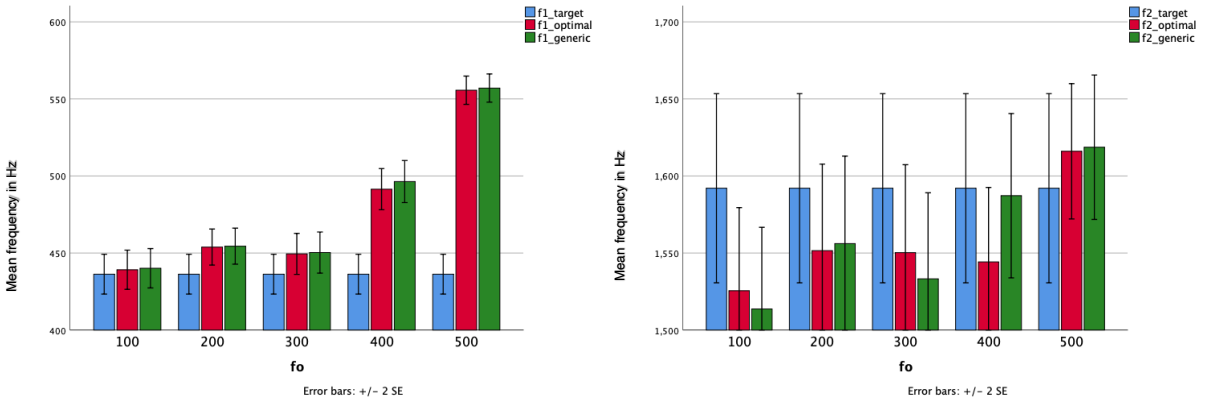


Figure 10. Mean estimations of F_1 (left) and F_2 (right) in hertz with both procedures compared to target at separate fo's. Scales are different and do not start from 0, but are comparable in terms of range. Error bars represent standard error, +/- 2 SE.

In Figure 11, estimations of F_1, F_2 with both procedures and their distance to target values for separate vowel types are shown. To relate this to thinking in terms of vowels, the F_2 is mostly correlated to frontness/backness of the vowel. Here, [æ], [ø] and [ʉ] have lower error and they are all front vowels, but so are [y], [i] and [e] and they have high error which makes it look as if frontness/backness of the vowel is not the explanation for the differences in correctness in estimations. Also [ɒ] has lower error and it is a back vowel, but so is [u] that has higher error and [o] that is in between [ɒ] and [u] in error. This result does not show a clear pattern, but it does paint the picture that the closeness of the vowel (i.e., when F_1 is low) affects the correctness in estimations, which in turn strengthens the assumption that it is the cases when f_o is so high that F_1 *cannot* be estimated correctly that affects correctness. Figure 11 also reveals that both procedures seems to overestimate F_1 overall, meaning estimations are higher in mean than target mean (for all vowels), but again, bear in mind that this could be because of the high f_o which in turn could mean that it is the actual f_o that is being mistaken for F_1. A pattern is revealed showing that F_2 seems overestimated for all back vowels ([ɒ], [o] and [u]) and underestimated for all front vowels ([æ] [e] [i] [ø] [ʉ] [y]), although for some vowels the difference relative to target is smaller than for others. The two estimation procedures do not show clear differences, though, they follow a very similar pattern overall. This was also seen in the ANOVA, since the interaction effect between formant estimation procedure and vowel type was not significant.

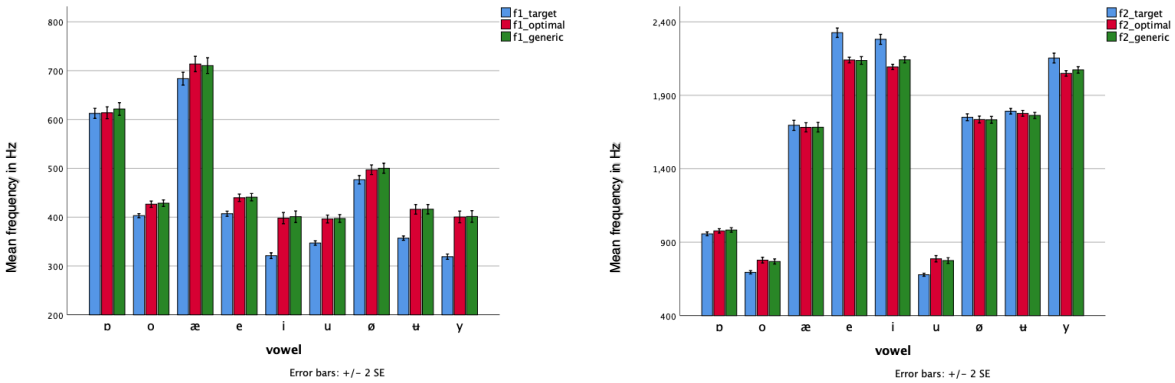
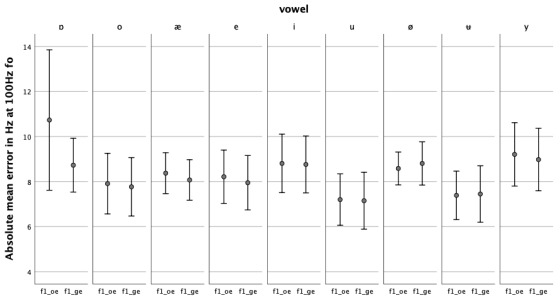


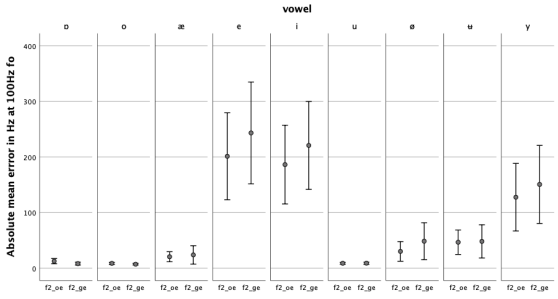
Figure 11. Mean values of target and estimations of F_1 (left) and F_2 (right) in hertz with optimal and generic ceilings for separate vowel types. Scales are different and do not start from 0, but are comparable in terms of range. Error bars represent standard error, +/- 2 SE.

3:4 Looking more closely at estimations of separate vowel types at separate f_o's

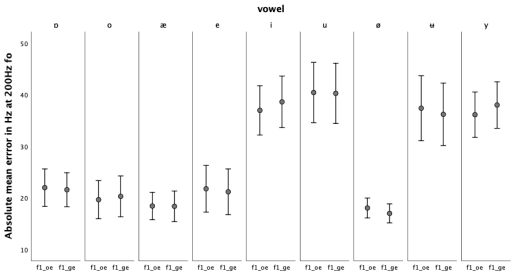
Looking more closely at estimation correctness in separate vowel tokens (with both procedures) at separate f_o levels did not reveal any clear pattern. None of the error bar diagrams in this section are log-transformed and the confidence intervals are likely to hide skewness in data. But with this in mind, these error bars are still illustrative, showing differences in correctness of estimations between procedures and f_o's and between some vowels which will be mentioned below. None of the procedures appears to be better in any specific condition (see Figure 12). The biggest difference between the procedures can be seen in F_2 at 400 Hz of f_o, where the optimal procedure shows more error than the generic procedure for the vowel type [i], [e], and [y] and at 500 Hz of f_o, where the optimal procedure shows more error than generic for the vowel types [i] and [u]. But at 100, 200 and 300 Hz of f_o it is the generic procedure that shows more error in F_2 for [e] and [i] and at 100 Hz also for [y]. These estimations of F_2 also show large variation in error. For F_1 the biggest difference between procedures can be seen at 300 Hz of f_o, where the vowel type [æ] has high error with very large variation in estimations with the generic procedure, but not with the optimal procedure. Otherwise, estimations with both procedures are very alike both in terms of amount of error and variation in F_1.



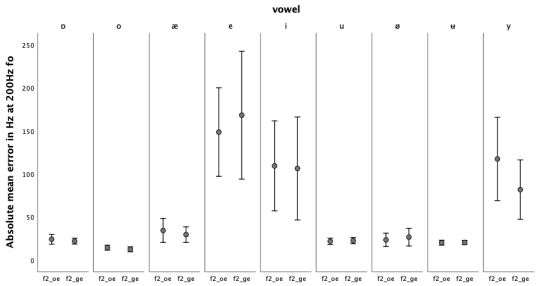
Correctness in estimations of F_1 at 100 Hz of f_o



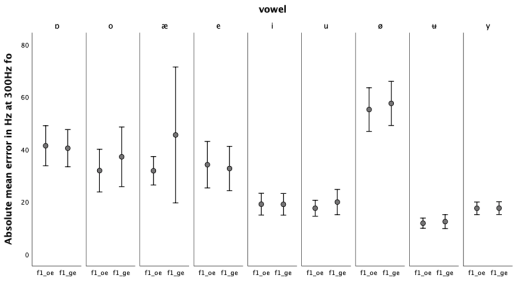
Correctness in estimations of F_2 at 100 Hz of f_o



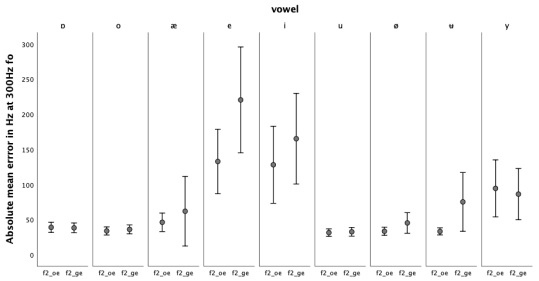
Correctness in estimations of F_1 at 200 Hz of f_o



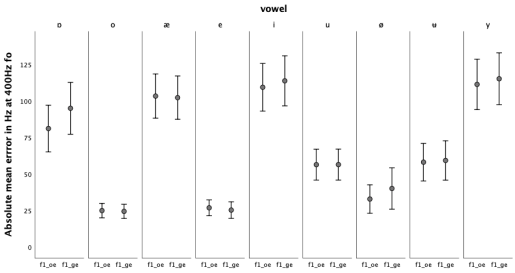
Correctness in estimations of F_2 at 200 Hz of f_o



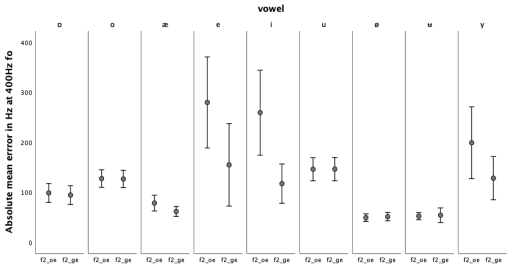
Correctness in estimations of F_1 at 300 Hz of f_o



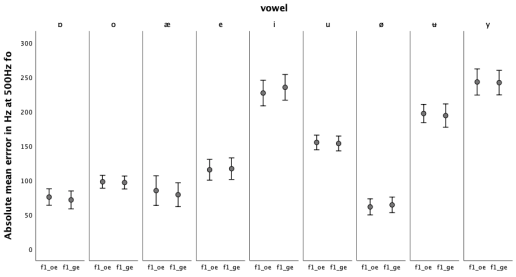
Correctness in estimations of F_2 at 300 Hz of f_o



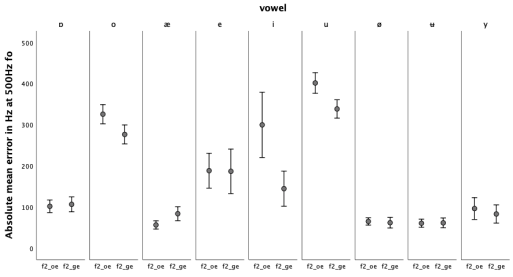
Correctness in estimations of F_1 at 400 Hz of f_o



Correctness in estimations of F_2 at 400 Hz of f_o



Correctness in estimations of F_1 at 500 Hz of f_o



Correctness in estimations of F_2 at 500 Hz of f_o

Figure 12. Correctness in estimations of F_1 (left) and F_2 (right) made with optimal ceiling (f1_oe, f2_oe) and generic ceiling (f1_ge, f2_ge), for each vowel type at all f_o levels separately; 100 Hz fo at the top, followed by 200 Hz f_o under and so on. Bars represent the 95% confidence interval.

3:5 Looking more closely at the distribution of ceilings

A check of the distribution of ceiling values showed that generic ceilings are organized just as they should be, with the ceiling at 5500 Hz for all female tokens and at 5000 Hz for all male tokens. The optimal formant ceilings vary from 4000 Hz up to 6500 Hz, but no clear distribution over vowel type or f_0 could be detected (see Figure 13). Some variation in the optimal ceiling across vowels was expected with higher ceiling values for front vowels (see Figure 3 in Escudero et al., 2009), however this doesn't seem to be the case here. As seen in Figure 14a, [i] and [u] have basically the same ceiling values.

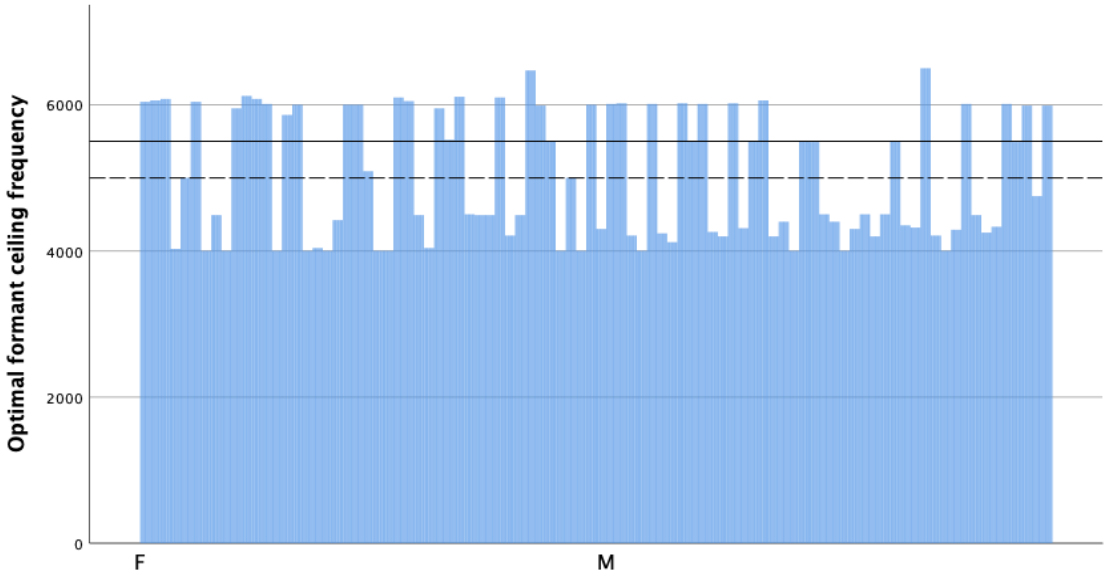


Figure 13. Distribution of optimal ceilings for all tokens. Female tokens to the left, male to the right. Tokens are represented in vowel type order: [ɒ] [o] [æ] [e] [i] [u] [ø] [ʉ] [y], each represented by token variations and f_0's 100-500 Hz for each type. Reference line added to show generic ceiling level at 5500 Hz (solid) for female tokens (f) and at 5000 Hz (dashed) for male tokens (m).

A closer look was taken at the distribution of optimal ceilings compared with generic ceilings for different vowel types. Figure 14a reveals that the optimal formant ceiling is lower than the generic ceiling in most cases. The largest difference in ceilings between the generic formant ceiling procedure and the optimal formant ceiling procedure is in the vowels [e i u ø], where the optimal ceiling is under the lowest generic ceiling (of 5000 Hz). Figure 14b shows that the optimal ceiling is lower with higher f_0, except for in the highest f_0 of 500 Hz. The optimal ceiling is also lower than the generic ceiling except for at the lowest f_0. Figure 14c shows that female tokens have higher ceilings than male tokens in both procedures, as can be expected. It also shows that generic ceilings are higher than optimal ceilings and that the difference between female and male ceilings seems to be greater with generic ceilings than with optimal ceilings.

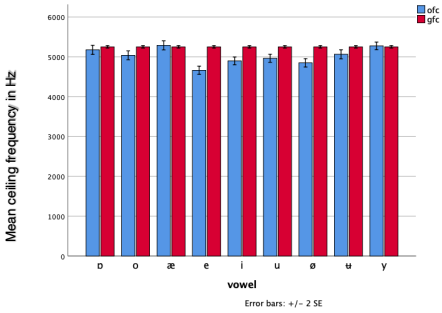


Figure 14a. Distribution of ceilings per vowel type. Error bars represent standard error, +/- 2 SE.

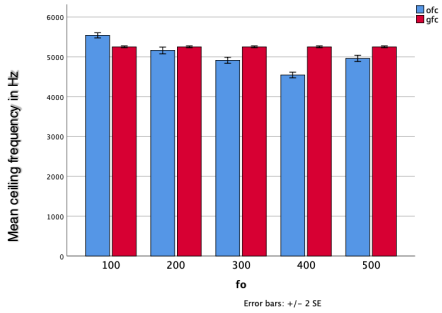


Figure 14b. Distribution of ceilings per f_o. Error bars represent standard error, +/- 2 SE.

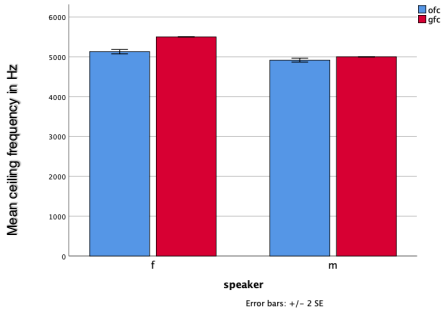


Figure 14c. Distribution of ceilings per speaker sex. Error bars represent standard error, +/- 2 SE.

3:6 Looking more closely at the problem of high f_o

Vowels with low F_1 (i.e., closed vowels) have more errors with a higher f_o. This can clearly be seen in Figure 15a, where target F_1 is shown next to estimations with both procedures at each f_o. At f_o of 400 and 500 Hz, all vowels are being overestimated, except for [ɒ]. Looking more closely at the data, this problem starts to occur already at f_o of 300 Hz, where vowel tokens that have F_1 under or close to 300 Hz are affected. These tokens are few, and individual tokens cannot be seen in this figure (as bars represent means) but these tokens are in vowel type [i] for male and some female tokens and in vowel types [u ʉ y], for all affected male tokens. Generally, F_1 is overestimated in all vowel types overall at 300 Hz, except for the open vowel type [æ]. Looking at f_o's at 400 and 500 Hz, this pattern is enhanced and also applies to all closed or half-closed vowels and both procedures. But it also varies somewhat between vowels, for example, at 400 Hz, the error for [e] and [o] is smaller than for [ɒ] and [æ]. This goes for both procedures; the optimal ceiling procedure does not perform better than the generic procedure.

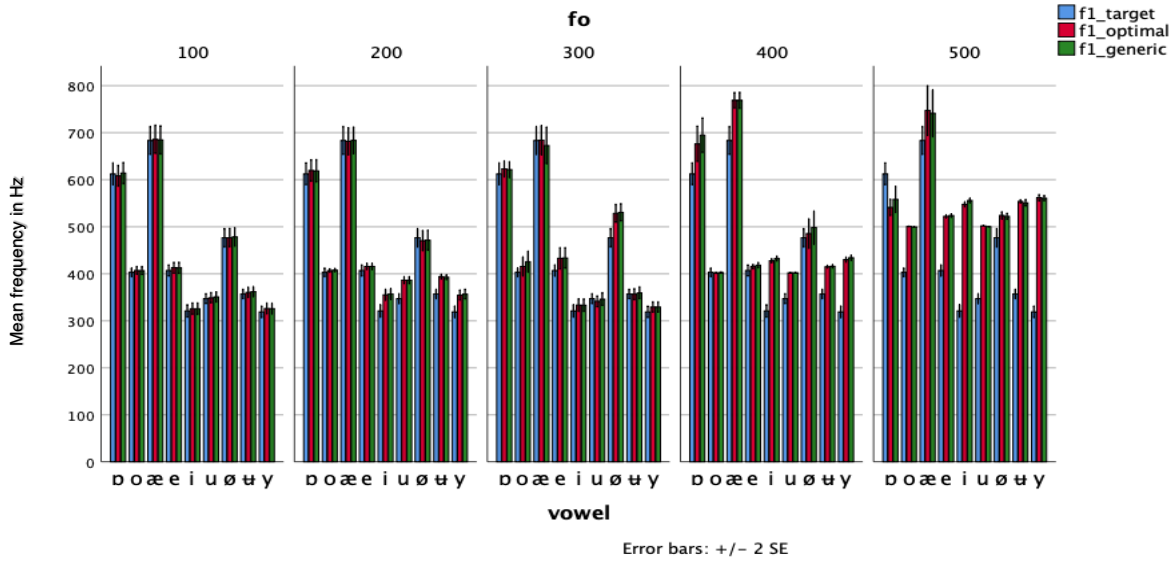


Figure 15a. Estimations of F_1 mean compared to target F_1 mean at all f_o levels. Error bars represent standard error, +/- 2 SE.

As for F_2, it can be seen in Figure 15b that F_2 is being overestimated for back vowels with rising f_o. This overestimation was seen already in Figure 11 for back vowels but here it is revealed that this occurs at higher f_o. The pattern observed earlier, with front vowels being underestimated as compared to the target, is also seen at every f_o here, but with some differences. For example, it can be seen that the generic procedure looks as if it is getting closer to the target than the optimal procedure in high f_o's of 400 and somewhat also at 500 Hz f_o in front vowels [e], [i] and [y]. Otherwise the procedures perform very similarly.

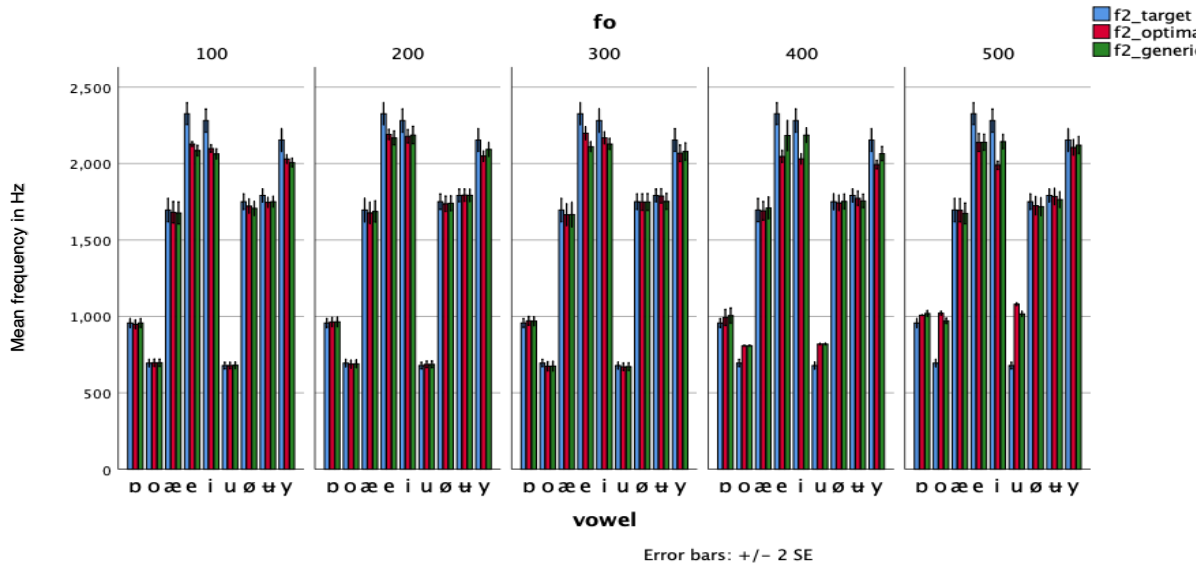


Figure 15b. Estimations of F_2 mean compared to target F_2 mean at all f_o levels. Error bars represent standard error, +/- 2 SE.

In Figure 16a, estimations of F_1 are shown compared to the target, divided into male and female for each f_o. Male closed vowels have lower F_1 values than female and are more be affected by high f_o. This can be seen in Figure 16a, as all closed or half-closed vowel types are more overestimated in male vowels (bottom) than in female vowels (top) at high f_o's. At the highest f_o, the open vowels seem to be underestimated in estimations of male vowel types and also in female [ɒ], but not for female [æ] or male [æ] with the generic procedure.

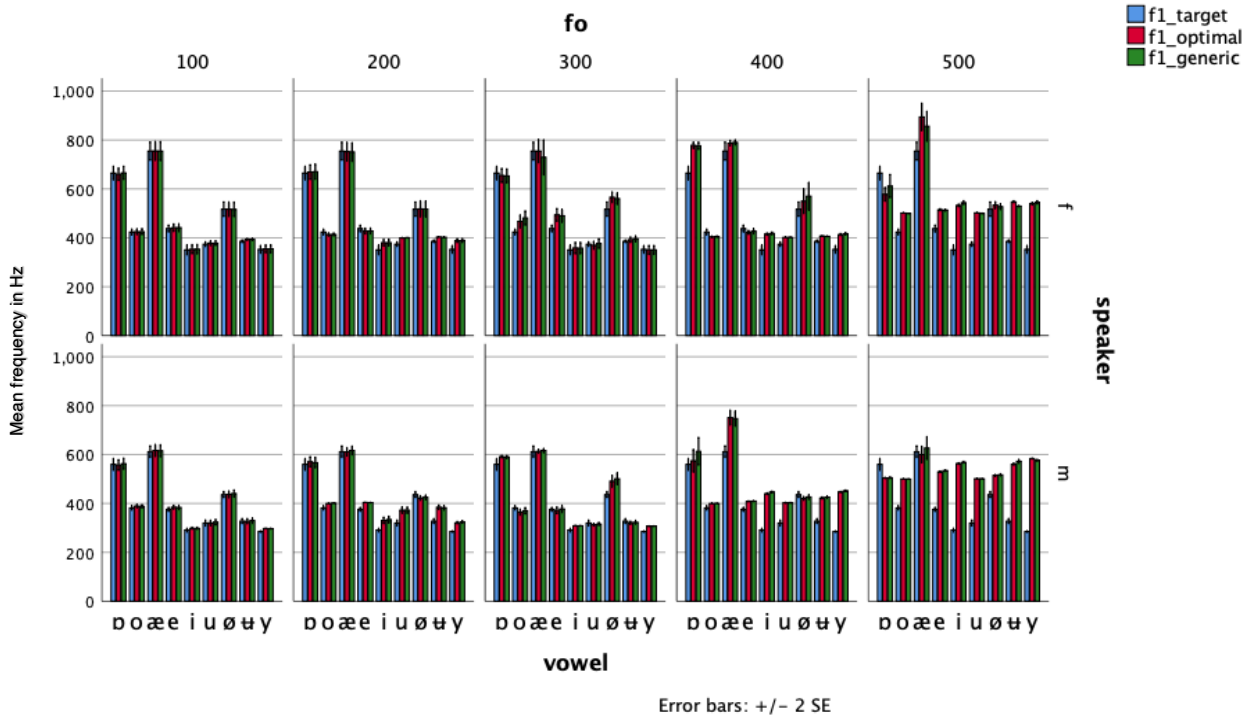


Figure 16a. Estimtions of F_1 mean compared to target F_1 mean at all f_0 levels, female and male tokens separately. Error bars represent standard error, +/- 2 SE.

In Figure 16b, estimations of F_2 are shown compared to the target, divided into male and female for each f_0. It can be seen that F_2 is underestimated by both procedures for female tokens but not as clearly for male tokens, which over all looks to be closer to the targets in general than female estimations, regardless of procedure.

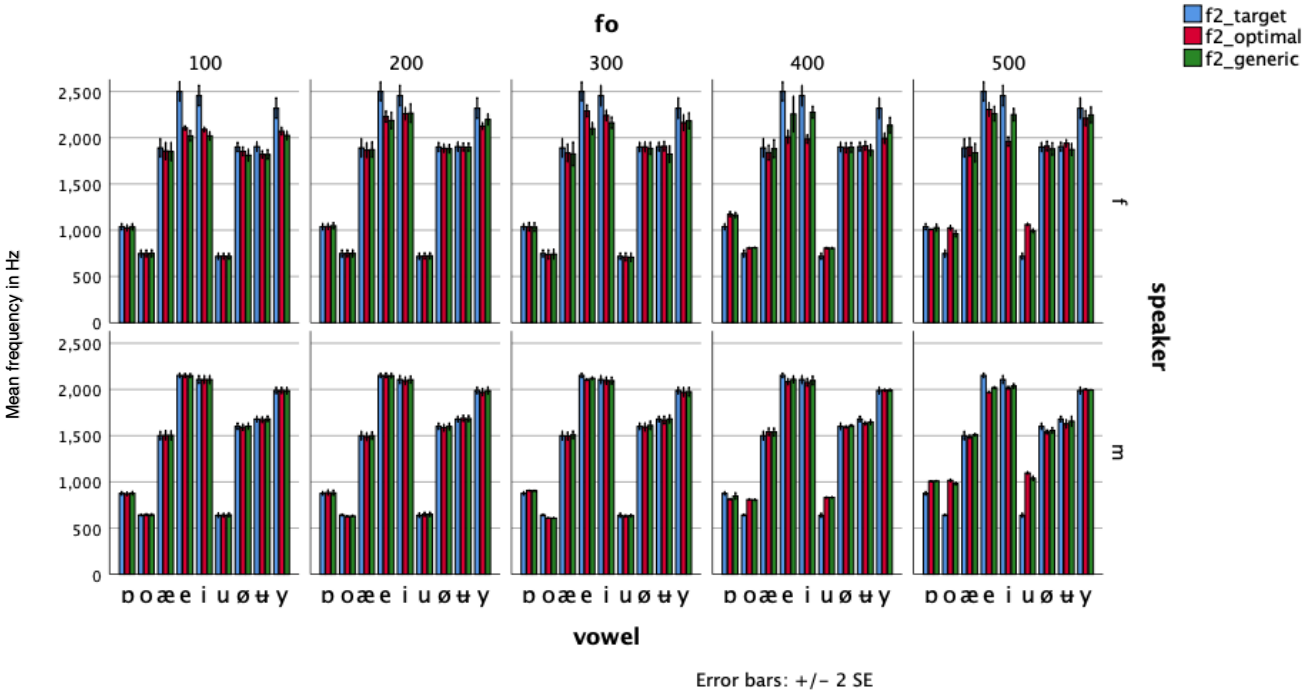


Figure 16b. Estimations of F_2 mean compared to target F_2 mean at all f_0 levels, female and male tokens separately. Error bars represent standard error, +/- 2 SE.

4 Discussion and conclusions

The optimal procedure was hypothesized to perform better than the generic procedure, but it did not. The main effect of formant estimation procedure indicated that the generic procedure performed somewhat better overall. The optimal procedure was also hypothesized to deal better with speaker sex, but did not, as no interaction effect between procedure and speaker sex was found. Although the data suggests that this may differ between formants, with the optimal procedure potentially performing somewhat better for female speakers in F_1 but not in F_2, this was not explicitly tested in the current study. As for dealing with high f_o, both estimation procedures perform less accurately with rising f_o overall, but an interaction effect between procedure and f_o indicates that the procedures are different with respect to impact of f_o. The generic procedure seems to perform better than the optimal procedure with rising f_o, except for at 300 Hz of f_o, where the optimal procedure performs slightly better. The optimal procedure does not handle variations, as in different vowels, better than the generic procedure. A main effect of vowel indicates that estimations of some vowels are more accurate than others. But the lack of interaction effect between estimation procedure and vowel type indicates that procedures do not perform differently for separate vowels, but rather both estimation procedures perform very similarly overall.

The high f_o and the closeness between F_1 and f_o (and the fact that F_1 is often even lower than f_o, "hiding" F_1) is likely a major contributing factor to the significant main effect of vowel. The difference in variation of formant frequencies in tokens is the reason that different vowel types are not being estimated with the same accuracy. Vowels with low F_1 (i.e., closed vowels) show more errors with higher f_o. This is not surprising, as when f_o is higher than F_1, there are no harmonics to represent the F_1. Correctness in estimations of the material in this study is already affected at 300 Hz f_o for some tokens. But this is expected, since the closed vowel tokens [i], [u], [ʉ], and [y] have low F_1 (with a frequency at its lowest F_1 in token variation of vowel type [i] at 267 Hz) and all tokens are represented with an f_o from 100 to 500 Hz. Therefore, the estimations of these vowels have more error already at an f_o of 300 Hz and the error becomes larger when f_o rises further. Problems can also be expected with F_2 when f_o is high, (because of sparseness in the spectrum it is likely to affect estimations of F_2 as well), but at least F_2 is never lower than f_o in frequency.

The problem with formant estimations when f_o is high seems inevitable, but in this data set this problem should at least be predictable by calculating the least expected error. The least expected error due to the high f_o, must be the difference between the target F_1 and the f_o (or 0 when F_1 is above f_o). This difference between F_1 target and f_o is calculated, giving a measure of the least error expected, when f_o is high. The mean values of the least expected error, for each f_o for all tokens, are shown in Figure 17 (named F1_fo_e), together with mean estimation errors from Monsen and Engebretsons study and from the present study (as in Figure 6). Showing them together in the same figure reveals how errors will always rise with rising f_o (when F_1 is lower than f_o), though problems with sparseness in harmonics still remain. This also shows clearly how all estimation methods will be affected. If the errors caused by rising f_o were to be subtracted from the values of the estimations, this would still result in error rising with f_o in this figure, meaning that the estimation methods may be as good as they could be under the circumstances. Note also that this is the least expected error in difference between f_o and F_1. Errors from F_2 are not included in this measure, but all estimation errors in the same figure are based upon estimations of both F_1 and F_2.

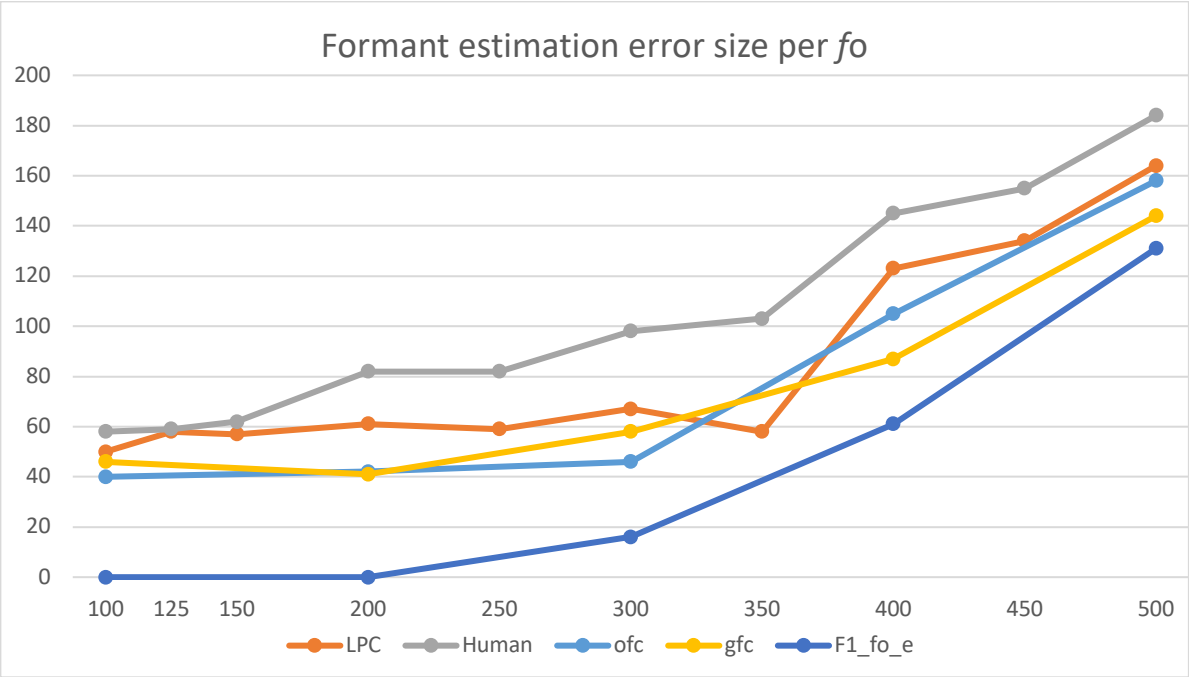


Figure 17. Formant estimations from previous and current study (as in Figure 7) with least expected error added for each f_0 from the current study in distance between F_1 and f_0 (F1_fo_e).

It seems clear that the optimal procedure has problems with high f_0 and does not perform better than the generic procedure, either at low or high f_0's. Estimating formants in material with high f_0 seems to be problematic for all methods mentioned in the background, except for inverse filter techniques, which seems to handle the high f_0, albeit often at the cost of a heavy manual workload. Only the automated inversed filter technique (IFC) remains to handle this problem (Watanabe, 2001).

An additional aspect that is important to put forward is that closed and half-closed vowels comprise seven out of the nine vowel types in this study, and open vowels only two, which could also affect the results, especially when looking at all vowels together. Furthermore, Monsen and Engebretson, (1983) used formant frequencies from English vowels and Escudero et al. (2009) looked at two dialects of Portuguese. Since the current study uses different vowels than those used by Monsen and Engebretson, and LPC performs differently on different vowels, a direct comparison cannot be made. Since the current study has so many closed vowel types with low F_1, it could make Monsen and Engebretsons' estimations look more correct when directly compared to estimations from the current study. But instead it looks as if the estimations in the current study are more correct than Monsen and Engebretsons' estimations.

In future studies, it would be interesting to take a closer look at how the optimal formant procedure performs compared to the generic procedure in the region of 300-400 Hz f_0. This could be done with the current procedure by adding more variations of tokens in smaller steps in variation of f_0, between 300-400 Hz, to see where the decrease in correctness starts and where estimations gets abruptly worse. This should also reveal whether the current

procedures follow the error pattern shown in Monsen and Engebretson (1983) at f_o of 350Hz (although this decrease in error could be too small to actually be significant) and also shed light on why the generic and optimal procedures change places in correctness at this f_o.

It would also be interesting to look more closely at data from estimations and analyze the selection of optimal ceiling. This would reveal whether the best estimations actually show the least combined variance, for F_1 and F_2. The distribution of optimal formant ceilings could be a reason why estimations are the way they are in the current study.

To summarize, the evaluation of the optimized formant ceiling procedure performed in this study revealed that it does not perform better formant estimations than the common procedure with generic settings. Neither does it handle variation between speaker sex better, although some differences where observed suggesting this may differ between separate formants. Both estimation procedures perform less accurately with rising f_o overall, but an interaction effect between procedure and f_o indicates that the procedures are different with respect to impact of f_o. The optimal procedure does not handle variations, as in different vowels, better than the generic procedure. Some vowels are more accurately estimated than others, but both estimation procedures perform very similarly overall. Taken together, the similarity between procedures and the unexpected lower performance by the optimal procedure is the overall picture of the evaluation and comparison performed in this study.

Acknowledgements

I would like to acknowledge the FORCE-project (Magnus Bergvalls Stiftelse 2018-02869; PI Lisa Gustavsson) at Stockholm University, for making this study possible. Furthermore, thanks to Ellen Marklund for the original idea to this study and together with Lisa Gustavsson for help, guidance and discussions during the whole process leading to insights, learnings and improvements of the study. Also, Ambika Kirkland contributed with valuable proofreading.

References

Acero, A. (1999). Formant analysis and synthesis using hidden Markov models. In *Sixth European Conference on Speech Communication and Technology*.

Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *The journal of the acoustical society of America*, *50*(2B), 637-655.

Bailey, P. J. (1983). Hearing for speech: The information transmitted in normal and impaired hearing. In *Hearing Science and Hearing Disorders* (pp. 1-34). Academic Press.

Broesch, T. L., & Bryant, G. A. (2015). Prosody in infant-directed speech is similar across western and traditional cultures. *Journal of Cognition and Development*, *16*(1), 31-43.

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer [Computer program]. Version 6.0.37 (February 3, 2018) to 6.1.09 (January 27, 2020). Retrieved from http://www.praat.org/

Chiba, T. & Kajiyama, M. (1941). *The Vowel: its nature and structure,* Kaiseikan, Tokyo.

Davis, B. L., & Lindblom, B. (2000). Phonetic variability in baby talk and development of vowel categories. Chapter in Emerging cognitive abilities in infancy, Cambridge University Press, Cambridge, Lacerda, von Hofsten, Heineman (eds.)

Deme, A. (2014). Formant strategies of professional female singers at high fundamental frequencies. In *Proceedings of the 10th International Seminar on Speech Production (ISSP)* (pp. 90-93).

Dissen, Y., Goldberger, J., & Keshet, J. (2019). Formant estimation and tracking: A deep learning approach. *The Journal of the Acoustical Society of America*, *145*(2), 642-653.

Eklund, I., & Traunmüller, H. (1997). Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. *Phonetica*, *54*(1), 1-21.

Escudero, P., Boersma, P., Rauber, A. S., & Bion, R. A. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *The Journal of the Acoustical Society of America*, *126*(3), 1379-1393.

Fant, G. (1960). Acoustic theory of speech production Mouton. *The Hague*.

Fant, G. (1970). *Acoustic theory of speech production* (No. 2). Walter de Gruyter.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of child language*, *16*(3), 477-501.

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental psychology*, *20*(1), 104.

Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., & Perkins, W. H. (1988). Relationship between changes in voice pitch and loudness. *Journal of Voice*, *2*(2), 118-126.

Granqvist, S. (2020). Sopran: Sound Processing and Analysis [Computer program]. Version 1.0.26 (August 14, 2020) Retrieved from http://www.tolvan.com/

Hertegard, S., & Gauffin, J. (1993). Voice source-vocal tract interaction during high-pitched female singing. In *Proceedings of the Stockholm music acoustics conference (SMAC)(Stockholm)(EJA Friberg, J. Iwarsson and J. Sundberg, eds.), Royal Swedish Academy of Music* (pp. 177-181).

Högberg, J. (1997). Prediction of formant frequencies from linear combinations of filterbank and cepstral coefficients. *KTH-STL Quarterly Progress Rep, Royal Inst. Technol. Stockholm, Sweden*, 41-49.

International Phonetic Association, & International Phonetic Association Staff. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Jessen, M., Koster, O., & Gfroerer, S. (2007). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech Language and the Law*, *12*(2), 174-213.

Kalashnikova, M., Carignan, C., & Burnham, D. (2017). The origins of babytalk: smiling, teaching or social convergence? *Royal Society open science*, *4*(8), 170306.

Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. *Infancy*, *4*(1), 85-110.

Ladefoged, P. (1962). Subglottal activity during speech. In *Proceedings of the Fourth International Congress of Phonetic Sciences* (Vol. 73, p. 91). Mouton The Hague.

Ladefoged, P., & Johnson, K. (1993). A Course in Phonetics, (International Edition).

Lindblom, B. E., & Sundberg, J. E. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *The Journal of the Acoustical Society of America*, *50*(4B), 1166-1179.

Marklund, E., & Gustavsson, L. (*in press*). A dynamic approach to vowel hyper- and hypoarticulation in infant-directed speech.

Monsen, R. B., & Engebretson, A. M. (1983). The accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction. *Journal of Speech, Language, and Hearing Research*, *26*(1), 89-97.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, *24*(2), 175-184.

Pfitzinger, H. R. (2003). Acoustic correlates of the IPA vowel diagram. In *Proc. of the XVth Int. Congress of Phonetic Sciences* (Vol. 2, pp. 1441-1444).

Pickett, J. M. (1980). *The sounds of speech communication: A primer of acoustic phonetics and speech perception*. Univ Park Press.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Sjölander, K., & Beskow, J. (2000). Wavesurfer-an open source speech tool. In *Sixth International Conference on Spoken Language Processing*.

Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30). MIT press.

Stevens, K. N., & House, A. S. (1955). Development of a quantitative description of vowel articulation. *The Journal of the Acoustical Society of America*, *27*(3), 484-493.

Sundberg, J., Lã, F. M., & Gill, B. P. (2013). Formant tuning strategies in professional male opera singers. *Journal of Voice*, *27*(3), 278-288.

Sundberg, J. (2001). *Röstlära: fakta om rösten i tal och sång*. Proprius.

Sundberg, J., & Skoog, J. (1995). Jaw opening, vowel and pitch. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, *36*(2-3), 43-50.

Sundberg, J. (1975). Formant technique in a professional female singer. *Acta Acustica united with Acustica*, *32*(2), 89-96.

Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., ... & Kreiman, J. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *The Journal of the Acoustical Society of America*, *137*(5), 3005-3007.

Titze, I. R., & Sundberg, J. (1992). Vocal intensity in speakers and singers. *the Journal of the Acoustical Society of America*, *91*(5), 2936-2946.

Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological science*, *11*(3), 188-195.

Traunmüller, H., & Eriksson, A. (1997). A method of measuring formant frequencies at high fundamental frequencies. In *Fifth European Conference on Speech Communication and Technology*.

Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. *Speech communication*, *49*(1), 2-7.

Wang, Y., Seidl, A., & Cristia, A. (2015). Acoustic-phonetic differences between infant-and adult-directed speech: the role of stress and utterance position. *Journal of child language*, *42*(4), 821-842.

Watanabe, A. (2001). Formant estimation method using inverse-filter control. *IEEE Transactions on Speech and Audio Processing*, *9*(4), 317-326.

Wood, S. (1989). The precision of formant frequency measurement from spectrograms and by linear prediction. *Speech Transmission Laboratory Quarterly Progress Report*, 91-93.

Xia, K., & Espy-Wilson, C. (2000). A new strategy of formant tracking based on dynamic programming. In *Sixth International Conference on Spoken Language Processing*.

APPENDIX

Appendix is showing results table from the ANOVA's Tests of within subjects Effects from SPSS.

Tests of Within-Subjects Effects Measure: MEASURE_1

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
ceiling	.235	1	.235	6.474	.011
ceiling * speaker	.002	1	.002	.055	.815
ceiling * vowel	.512	8	.064	1.767	.079
ceiling * fo	1.583	4	.396	10.926	.000
ceiling * speaker * vowel	1.864	8	.233	6.431	.000
ceiling * speaker * fo	1.048	4	.262	7.230	.000
ceiling * vowel * fo	4.407	32	.138	3.801	.000
ceiling * speaker * vowel * fo	6.069	32	.190	5.235	.000
Error(ceiling)	78.253	2160	.036		
formant	59.084	1	59.084	201.497	.000
formant * speaker	82.927	1	82.927	282.810	.000
formant * vowel	73.143	8	9.143	31.181	.000
formant * fo	41.637	4	10.409	35.500	.000
formant * speaker * vowel	96.470	8	12.059	41.125	.000
formant * speaker * fo	24.632	4	6.158	21.001	.000
formant * vowel * fo	154.347	32	4.823	16.449	.000
formant * speaker * vowel * fo	44.984	32	1.406	4.794	.000
Error(formant)	633.364	2160	.293		
ceiling * formant	.219	1	.219	5.758	.016
ceiling * formant * speaker	.027	1	.027	.707	.400
ceiling * formant * vowel	.627	8	.078	2.056	.037
ceiling * formant * fo	1.839	4	.460	12.061	.000
ceiling * formant * speaker * vowel	1.596	8	.200	5.235	.000
ceiling * formant * speaker * fo	.186	4	.046	1.220	.300
ceiling * formant * vowel * fo	4.653	32	.145	3.815	.000
ceiling * formant * speaker * vowel * fo	5.511	32	.172	4.519	.000
Error(ceiling*formant)	82.330	2160	.038		

Stockholm University
SE-106 91 Stockholm
Phone: 08 – 16 20 00
www.su.se

Stockholms universitet