Formants

Daniel Aalto

Department of Communication Sciences and Disorders,

Faculty of Rehabilitation Medicine, University of Alberta, Canada;

Institute for Reconstructive Sciences in Medicine,

Misericordia Community Hospital, Edmonton, Canada

Jarmo Malinen

Department of Mathematics and Systems Analysis,

Aalto University, Finland

Martti Vainio

Department of Modern Languages, University of Helsinki, Finland

Abstract

Formant frequencies are the positions of the local maxima of the power spectral envelope of a sound signal. They arise from acoustic resonances of the vocal tract air column, and they serve to differentiate mainly the vowels. In running speech they are crucial in signalling the movements with respect to place of articulation. Formants are normally defined as accumulations of acoustic energy estimated from the spectral envelope of a signal. However, not all such peaks can be related to resonances in the vocal tract as they can be caused by the acoustic properties of the environment outside the vocal tract and sometimes resonances are not seen in the spectrum. Such formants are called spurious and latent, respectively. By analogy, spectral maxima of synthesised speech are called formants if they are similarly associated to resonant frequencies, related to poles of a digital filter contained by the synthesiser. Conversely, speech processing algorithms can detect formants in natural or synthetic speech by modelling its power spectral envelope using a digital filter. Such detection is most successful from male speech with a low fundamental frequency where many harmonic overtones coincide with any of the vocal tract resonances that lie at higher frequencies. For the same reason, reliable formant detection from high pitch female or children's speech is inherently difficult, and many algorithms fail to faithfully detect the formants corresponding to lowest vocal tract resonant frequencies.

Keywords: Formant, Resonance, Phonetics, Speech acoustics, Speech production

Formants

Introduction

Vibration and resonance are fundamental aspects of the physical world. Celestial systems have orbital resonances, and quantised energy levels of electrons in an atom can be understood as resonances as well. Resonant behaviour can be observed in mechanical and electrical systems such as pendulums, strings, tank circuits, and transmission lines as well as acoustical systems such as isolated air columns in wind instruments. Animals have adapted resonant phenomena, at least, for their locomotion and – which is more pertinent to the current context – for acoustic communication.

The air column in the human *vocal tract* (VT) serves as a primary example of an acoustic resonator whose resonant behaviour can be changed by external control actions. It is a system that can store and release vibrational energy according to physical principles described by acoustic equations. As a control system, the articulating vocal tract is fairly complex with several controllable degrees of freedom. As users of spoken language, we learn to control this biomechanical system through available control actions to a communicative end. We do this by turning our articulatory gestures into features of sound, thus sculpting the sound shape that we emit from our vocal organs.

Arguably, the most important communicative function of our vocal apparatus is the formation of spectral properties of continuous speech signal into features that are efficient in producing acoustic – and, therefore, perceptual – contrasts. In voiced speech, the vocal tract functions as a filter for the air flow pulses, produced in the larynx by the vibrating vocal folds. In this process the acoustic resonances in the vocal tract have sound shaping consequences that are called *formants*. Formants are thus – at least in principle – measurable features of the speech signal. However, both the definition of the concept of formant as well as their extraction from the speech signals is not as straightforward as one would hope.

In this chapter, we explore resonances of the human vocal tract, relate the acoustic resonances to formants, describe some formant extraction algorithms, and discuss some neighbouring concepts as well as the challenges related to both the definition of the term and the empirical estimation of their values.

Resonances of the vocal tract air column

Any air volume, bounded by a closed surface, has a discrete set of *resonant frequencies* and corresponding *(standing wave) modal patterns* of acoustic pressure variations. Examples of such surfaces are given in Figure 1 where the air/tissue interface of VT from a male test subject is shown while producing vowel sounds during a Magnetic Resonance Imaging (MRI) scan. As such, the air/tissue interface does not constitute a *closed* surface unless the openings at mouth and glottis are terminated by a virtual boundary condition interface, describing the interaction with the acoustic environment.

Given a surface geometry, the resonant frequencies and modal patterns of the interior air volume can be computed from a number of acoustic models, often described by Partial Differential Equations (PDE). A simple and common 3D acoustic resonance model is the classical *Helmholtz equation*. If a computationally lighter model for the longitudinal part of the acoustics for low frequencies is sufficient, a time-independent Webster's horn model of transmission line type can be used. Energy dissipation through tissue walls and boundary friction can be included in both the models, if deemed necessary. In the context of human vocal tract, the Webster's model usually is sufficient in treating the three lowest resonant frequencies that lie under 4 kHz at least in adult males. Both of these resonance models are *eigenvalue problem* -versions of corresponding time-variant PDEs: the wave equation and (time-dependent) Webster's equation, respectively, as treated in, e.g., A. Aalto, Lukkari, and Malinen (2015). For the rest of this article, the vocal tract resonances refer to those computed from a Finite Element (FE) discretisation of the Helmholtz equation as described in (Hannukainen, Lukkari, Malinen, & Palo, 2007, Section 2), and they are denoted by $f_{R_1}, f_{R_2}, \ldots$ and always enumerated in the increasing order. We point out that neither the resonant frequencies nor the modal patterns can be given exact descriptions by mathematical formulas in domains as complicated as the vocal tract. Thus, numerical approximation and computational modelling are necessarily required.

Considering a tube with uniform diameter and finite length whose one end is

closed, the lowest resonant frequency differs by one octave depending whether the other end (corresponding to the mouth) is closed or (idealised) open. Similarly, the environment acoustics near the mouth opening has a very significant effect on each $f_{R_1}, f_{R_2}, \ldots$, and the effect is more pronounced when the mouth opening area is large. In physically realistic modelling, the radiation of the sound energy to the exterior space (containing, e.g., the test subject's face) must thus be taken into account. The most simple assumption – known as the *Dirichlet boundary condition* at the mouth – used in Hannukainen et al. (2007) must then be deemed as too inaccurate. Instead, the external space should be included in the computational model in the same way it appears in the measured speech whose spectral features the model attempts to replicate; see, e.g., Arnela, Guasch, and Alías (2013). For example, room resonances may be observed in the sound signal as *external formants*, or the environment may appear as infinitely large empty space (save for the test subject's head and vocal tract) in measurements carried out in an anechoic chamber. The VT resonances can also be measured via exciting the VT air column by an artificial sound source (see e.g. Epps, Smith, and Wolfe (1997)).

Formants

The resonances of the VT modify the amplitudes of the harmonics of the glottal source signal. At the beginning of 1890s, the concept of the *formant* emerged to describe the broad peaks in the frequency spectrum of vowels caused by the VT resonances. For historical accounts of formant analysis, see Fulop (2011) and Titze et al. (2015). The modern definition of the formant relies only on the sound signal itself even though the usual classification of formant frequencies necessarily requires some sort of *a priori* information about the system that produced the signal. However, always requiring an explicit, tractable, mechanistic link from VT resonances to formants suffers from cumbersome practical limitations one cannot afford.

Treating formants purely as spectral peaks raises a question of what kind of spectral peaks should count as formants. A vowel produced with a low, constant fundamental frequency creates a comb-like spectrum where every harmonic has a

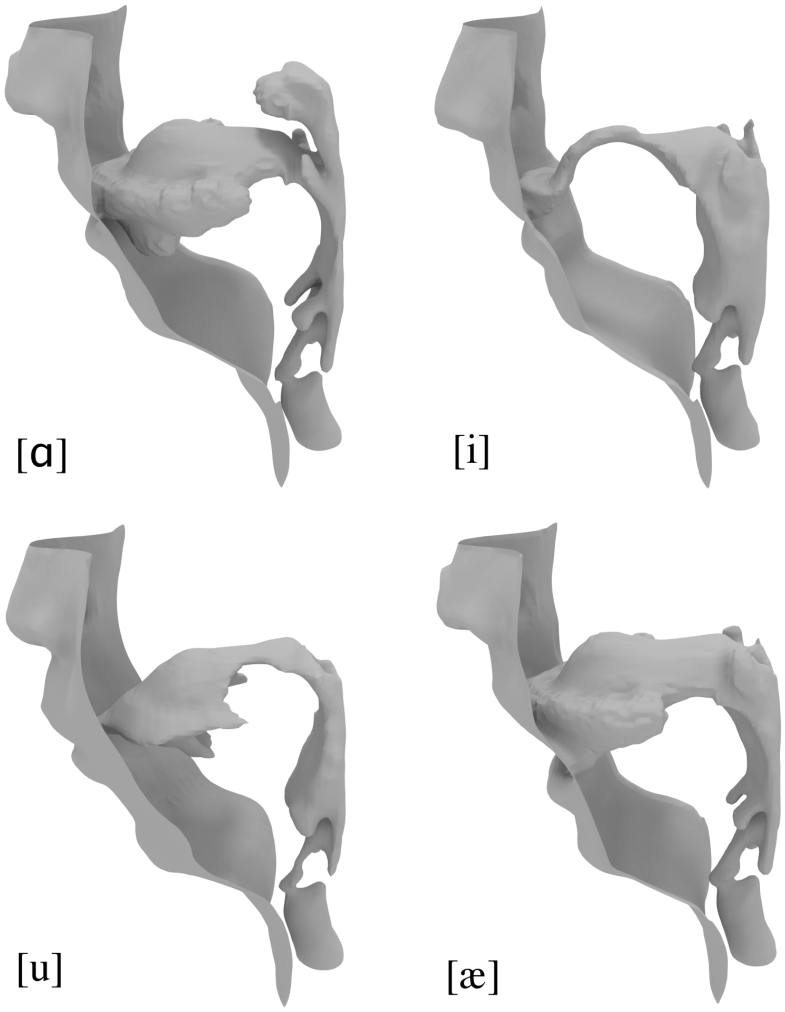

Figure 1. Anatomic air/tissue interfaces corresponding to Finnish vowels [ɑ, i, u, æ] uttered by a male test subject. The models have been extracted from Magnetic Resonance Images (MRI). Note that the velopharyngeal port is open in [ɑ].

spectral peak of its own. However, these source generated peaks are not considered as formants. Instead, a broader peak created by multiple harmonics could count as one. Formally, these broader peaks can be identified as the peaks of the *power spectral envelope* of the signal as specified below. The location and the breadth of the peak in the spectral envelope are called *formant frequency* and *formant bandwidth*, respectively. However, the power spectral envelope is not uniquely defined but typically depends on the extraction method and its parametrization (including the number of parameters used). In selecting the suitable parameters, some *a priori* information is required to

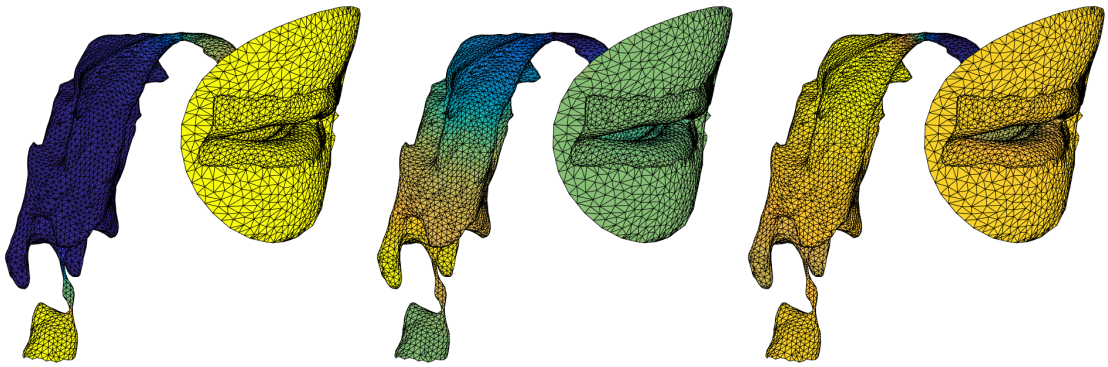

Figure 2. The pressure distributions of three lowest resonant modes at frequencies 206 Hz, 2540 Hz, and 2949 Hz computed from a Helmholtz model by Finite Element Method. The anatomic geometry corresponds to the Finnish vowel [i] uttered by the test subject shown in Figure 1. There is an artificially joined surface surrounding the lips used for matching the suitable exterior boundary condition required by the model.

make an informed decision. The use of *a priori* information is motivated by the desire of having a good match between resonant frequencies of the VT and formant frequencies. This relationship will be explored in more detail later in this chapter.

A signal-based definition of a formant states that

> a formant is a relatively broad peak in the power spectral envelope of a quasi-stationary vowel-like signal of sufficient length.

Some reservations and explanations are now in order. In *vowel-like* signals of practically constant fundamental frequency f_o (by quasi-stationarity), the upper harmonics kf_o, $k = 2, 3, \ldots$, are regarded as narrow spectral peaks, and the envelope peaks understood as formants are required to be significantly broader compared to them. Secondly, formants cannot be reliably extracted or even defined in very short samples. One full period of vocal fold oscillation is required for reasonable spectral analysis but the resolution improves with longer signals.

The above signal-based definition (or its variants) is broadly but not the only universally accepted definition. Sometimes formants are used to mean the resonances of the VT which can be confusing if the formants are used in their above given meaning as

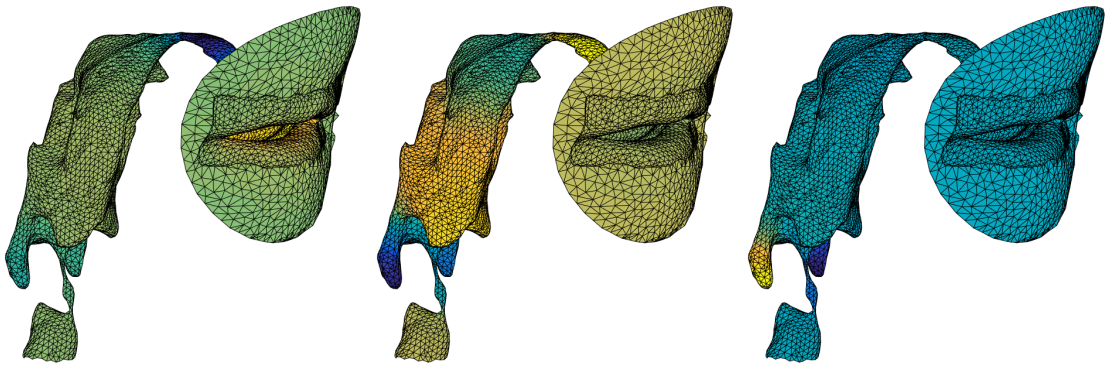

Figure 3. The pressure distributions of the fourth, fifth, and sixth lowest resonant modes at frequencies 3756 Hz, 4003 Hz, and 5308 Hz of the Finnish vowel [i]. The anatomic geometry and computational details are same as in Figure 2.

well. At times, formants refer to specific parameters in a power spectral envelope or a VT filter model (e.g., pole positions in an all-pole model). Finally, formants can refer to the local activity maxima along the tonotopic axis at various stages of the ascending auditory pathway. In terms of frequencies, these definitions often give similar results for vowels. However, these are distinct concepts and an explicit definition of formant is recommended when used in a scientific context; see also Titze et al. (2015).

Power spectral density and its envelope

The power spectral envelope of a sound signal is the slowly varying component (in the frequency axis) of its power spectral density. As such, it is an approximation (albeit typically a poor one) for the spectral density. Its usefulness in formant analysis of natural speech essentially depends on the lack of detail which reveals the fingerprints of vocal tract resonances by downplaying the sharp harmonics of the fundamental frequency f_o. Spectral envelopes are often produced by spectral density estimation algorithms when they are intentionally configured to run in low resolution.

Methods for estimating the power spectral density from signals can be divided into two classes: *Parametric* and *non-parametric methods*. Parametric methods make an explicit assumption of a model that has produced the observed spectral density from an input, typically regarded as a noise process. For example, all Autoregressive Moving

Average (ARMA) models are parametric as they assume an undelying rational filter, and these models include the variants of Linear Predictive Coding (LPC or LP) such autocorrelation, covariance, and Burg's method; see, e.g., Stoica and Moses (2005). An example of a non-parametric method is the plain Fast Fourier Transform (FFT) applied on a sampled signal where the sampling rate is not understood as a "model parameter" though it, of course, is a parameter. In speech processing, parametric methods are used almost exclusively for estimating spectral densities and their envelopes.

The parametric approaches for computing the spectral envelope divide further into two classes:

1. ARMA-type "on-line" methods that dynamically produces a low-order rational model for the resonator, based on the already recorded part of the signal.

2. Smoothing by, e.g., direct rational approximation (such as the Padé approximation) of the pre-computed and stored high-resolution power spectral density or its spectral factor.

In both cases, a low-order (and, hence, a rather poor) rational approximant for the power spectral density is obtained, and it provides us with a version of the spectral envelope. The order (i.e., the number of the poles and possibly zeroes) of the rational approximant can and should be chosen reasonably, considering the nature of speech signals. It is usual to use all-pole models (e.g., autoregressive (AR) models) without zeroes since the impedance of an acoustic waveguide (lacking side cavities) does not have transmission zeroes at all; see Makhoul (1975), El-Jaroudi and Makhoul (1991). It should be noted that for vowels with high f_o, the formant extraction by conventional AR methods is more difficult since the harmonics kf_o, $k = 2, 3, \ldots$ are sparse within the peaks of, in particular, low vocal tract resonances.

Most autoregressive spectral estimation methods of speech optimise the model parameters by minimising a residual using the Least Squares Error criterion, computed typically over a time frame of $10 - 30\,\text{ms}$ spanning $1 - 10$ glottal cycles. In such methods, all the time instants of the speech waveform are treated with equal weight in the model optimisation. In Weighted Linear Prediction (WLP), however, a positive,

real-valued weight function enters the residual, and it is used to emphasise parts of the speech waveform differently. Such weighting alleviates the challenges of formant extraction from signals with high f_o; see Alku, Pohjalainen, Vainio, Laukkanen, and Story (2013) where Weighted Linear Prediction with Attenuated Main Excitation (WLP-AME) is introduced.

Peaks of the envelope and poles of the approximant

The peaks of the power spectral envelope have a correspondence to the poles of the rational approximant. In parametric methods, some *a priori* information is useful for deciding how many poles one should place to a given frequency interval. As *a priori* information one can use, for instance, the VT resonant frequencies associated to the VT configuration (as is done below for classification of the envelope peaks into the sequence of formant frequencies), or plainly the knowledge of the intended vowel combined with pre-tabulated reference values.

Parametric methods allow the user to choose the number of poles (i.e., the order of the AR model) for the power spectral envelope in one way or another. Placing too few poles near some frequency interval may miss or misplace some of the formants in their *a priori* expected positions. Recorded signals always contain acoustic signatures of the exterior space in contrast to the VT considered as the interior space. If the recordings for some reason are carried out in a physically very constrained environment, the exterior resonances may appear in the same frequency range as the lowest formants; see, e.g., D. Aalto et al. (2014), Kuortti, Malinen, and Ojalammi (2017). In these situations, one should use sufficiently high order parametric method not to merge into a single peak a true formant of the VT and a resonance of the exterior space.

Allowing too many poles will produce multiple peaks in the envelope for a single formant. Increasing the number of poles excessively will make the spectral envelope approximate well the original, high-resolution power spectral density of the signal, thus spoiling the sought after separation between the harmonics of the fundamental frequency and the contribution of the VT acoustic resonances. However, it may be

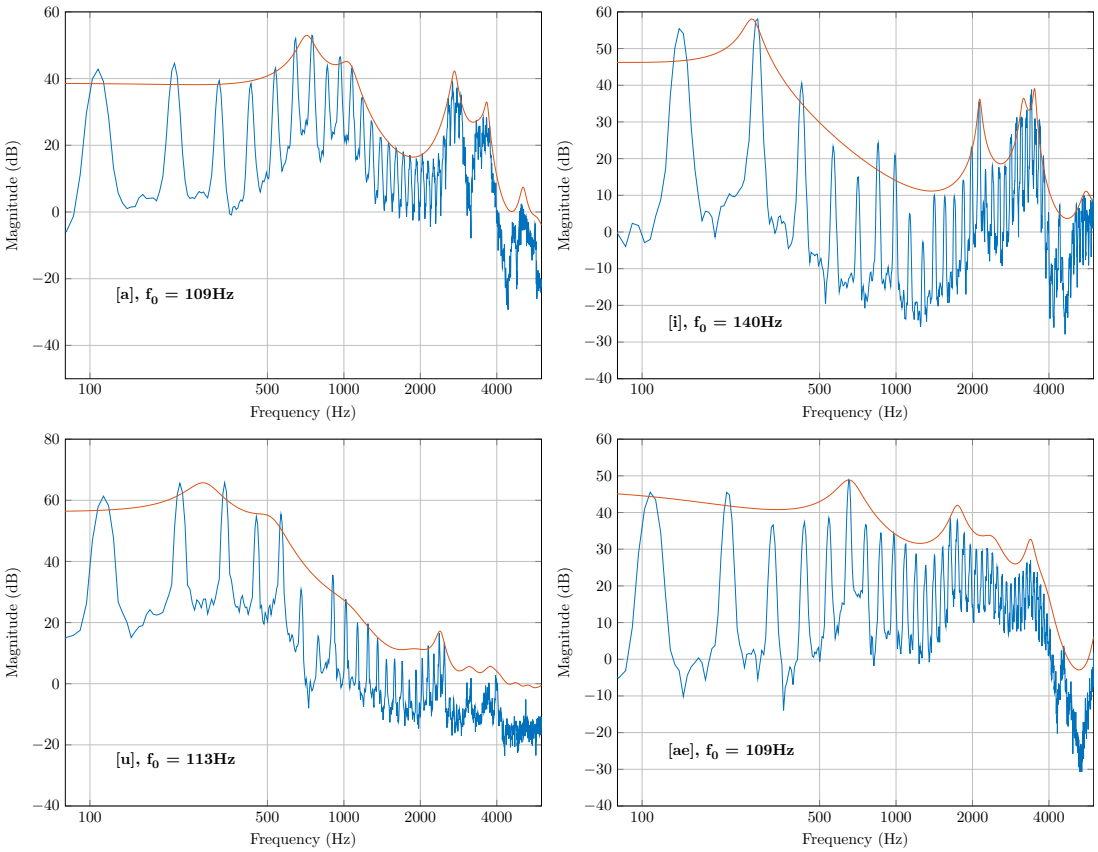

Figure 4. Power spectral densities and envelopes between 100 Hz . . . 5 kHz of Finnish vowels [ɑ, i, u, æ] pronounced by a male test subject. The spectral envelope has been raised to improve readability. Observe that a prominent spectral concentration exists right above 4 kHz in spectral densities of vowels [i, u, æ] that has not been detected in the spectral envelope.

desirable to have several poles corresponding to a peak of the spectral envelope that has too high bandwidth to be sufficiently approximated by a single pole. For this reason, it may be more practical to use the positions of the peaks or the centroids of peak clusters (determined by suitable means depending on the application) of the spectral envelope for defining formants instead of directly using the pole positions in the complex plane.

Examples of vowel spectra

Figure 4 shows power spectral densities and their envelopes of Finnish vowels [ɑ, i, u, æ] uttered by the test subject whose VT configurations are shown in Figure 1. The

recordings have been carried out in an anechoic chamber with a microphone placed in front of the mouth. The sample rate of the recoding was 44.1 kHz, the analysed portion of the signal is 500 ms from the middle of a sustained vowel production, with 8192 point FFT to describe the power spectral density. Burg's method (see Burg (1975)) has been used (with its order chosen by trial and error) to extract the power spectral envelopes suitable for the extraction of formant frequencies and bandwidths.

The relatively broad formant peaks of the spectral envelope can easily be distinguished from the comb-like harmonic spectral structure of the glottal source in Figure 4. The peak centre frequencies of the formants are denoted by $f_{P_1}, f_{P_2}, \ldots$, and they are always enumerated in the increasing order the same way as the vocal tract resonances. The numerical values for $f_{P_1}, f_{P_2}, \ldots$ from Figure 4 are given in Table 1. For ease of presentation, we use here the local maxima of the spectral envelope as peak centre frequencies, but one could have extracted almost the same frequencies directly from the pole positions of the rational filter, produced by Burg's method.

Table 1

The peak centre frequencies and their approximate 3 dB *bandwidths (in* Hz*) from spectral envelopes of Fig. 4. The peak centres correspond to the local maxima of the power spectral envelope. Those peaks were marked with an asterisk for which the bandwidth cannot be determined.*

Vowel	f_{P1}	f_{P2}	f_{P3}	f_{P4}	f_{P5}
[ɑ]	710 (130)	1028 (120)	2719 (160)	3650 (210)	5075 (300)
[i]	269 (30)	2143 (60)	3182 (190)	3510 (100)	5626 (670)
[u]	280 (150)	490 (*)	2353 (430)	3720 (900)	
[æ]	651 (160)	1739 (210)	2342 (*)	3391 (610)	

The relationship between formants and resonances

According to the source filter theory of speech production, the spectrally rich glottal signal is shaped (filtered) by the VT resonances. These resonances can be

controlled by adjusting the configuration of the supraglottal articulators; see MacDonald, Purcell, and Munhall (2011). In the following paragraphs, a constructive, detailed, and exhaustive method to define formants is proposed. In the absence of resonance information, other data (such as tabulated average formant frequency values of a homogenous speaker group) can be used as *a priori* information.

Extraction of formant frequencies from spectral envelopes

Considering the spectral envelope data given in Figure 4 and peak centre frequencies given in Table 1 for the formants, we need use some form of *a priori* information to conclude which of the frequencies $f_{P1}, f_{P2} \ldots$ actually are related to some of the vocal tract resonance frequencies $f_{R1}, f_{R2} \ldots$. We adopt the following definition:

> The peak centre frequency f_{P_j} for $j = 1, 2, \ldots$ is a *(vocal tract) formant frequency* if it can be associated to some of the acoustic resonances f_{R_k} with $k = 1, 2, \ldots$ of the same, or comparable, vocal tract anatomic configuration of the same vowel. In this case, we write $F_k = f_{P_j}$.

As a result of this process, one obtains a nondecreasing sequence of formant frequencies $F_1, F_2, \ldots$ that may have gaps (i.e., the value of F_k not defined for some k) or iterations (i.e., $F_k = F_{k+1}$ for some k). We earlier defined formants as peaks of the power spectrum, and the latter definition given above concerns VT formant frequencies. Usually, the word "formant" is used to refer to both of these concepts since the correct interpretation can be made from the context.

Obviously, the acoustic resonances f_{R_k} are not always available for formant estimation. If the speech measurement can be carried out in optimal conditions, and if the spectral envelope can be produced with exactly one peak centre frequency $f_{P_1}, f_{P_2} \ldots$ at each expected formant position (which expectation constitutes another kind of *a priori* information), then one is able to produce the formant frequency sequence $F_1, F_2, \ldots$ without knowing $f_{R_1}, f_{R_2} \ldots$ at all. Unfortunately, a number of issues may arise where the above given process, leading to values $F_1, F_2, \ldots$, must be refined:

1. If there are several peak centre frequencies f_{P_j} that are associated to the same

resonance f_{R_k}, then the formant frequency F_k is not uniquely defined. To get a value for F_k, either a lower order spectral envelope could be used, or the value could be defined as a cluster centroid of f_{P_j}s in question.

2. The centre peak frequency f_{P_j} in the power spectral envelope is a *spurious formant* if it does not coincide with any of the vocal tract formants $F_1, F_2, \ldots$ as defined above.

3. If vocal tract resonance frequency f_{R_k} for some k cannot be associated with any of the peak centre frequencies f_{P_j} for $j = 1, 2, \ldots$ of the power spectral envelope, then we say that F_k is a *latent formant* whose value is left undefined.

4. If the same formant frequency value F_k appears several times in the sequence $F_1, F_2, \ldots$, then F_k is a *compound formant*.

Spurious formants are artefacts of the power spectral envelope, essentially not due to the test subject's VT, that may result from, e.g., the frequency response of the speech measurement arrangament or the acoustics of the environment (hence, the concept of *external formants*). A latent formant serves as a place holder in the formant sequence $F_1, F_2, \ldots$ for a value that cannot be detected from the power spectral envelope under study even though it is known that a VT resonance must lie in a neighbourhood. Finally, it may be practical to remove the iterated versions of compound formants from the sequence $F_1, F_2, \ldots$ and re-index the sequence so as to obtain a strictly increasing sequence of formant frequencies with, of course, the exceptions of latent formants.

The lowest resonance of the VT in [i] configuration appears typically under 300 Hz, and it is likely to yield a latent F_1 if the fundamental frequency satisfies, e.g., $f_o > 600$ Hz in female phonation. The fourth formant F_4 may be a compound formant since it typically corresponds to a cluster of several VT resonances that often cannot be reliably spectrally separated in samples of natural speech. Also, the singer's formant is an example of a compound formant.

To indicate how these definitions and their refinements are used in practice, we next carry out the formant analysis of [i] based on the power spectral envelope shown in Figure 4 and the peak centre values in Table 1 using the vocal resonance data from

Figures 2—3 as the *a priori* information. We get $F_1 = 269\,\text{Hz}$, $F_2 = 2143\,\text{Hz}$, $F_3 = 3182\,\text{Hz}$, $F_4 = 3510\,\text{Hz}$, F_5 is latent with $f_{R_5} = 4003\,\text{Hz}$, and $F_6 = 5626\,\text{Hz}$. The fundamental frequency in this sample is $f_o = 140\,\text{Hz}$ which is low enough not to make F_1 latent. Observe that there is a discernible peak in the power spectral density of [i] in Figure 4 at the expected position of F_5 but it has not been captured in the spectral envelope. Had the two peaks corresponding to F_3 and F_4 in the spectral envelope of [i] in Figure 4 been merged together, then F_3 would have been a compound formant with its value between 3182 Hz and 3510 Hz, F_4 latent, and $F_5 = 5626\,\text{Hz}$. There are no spurious formants in [i] in Figure 4 since the recording has been carried out in optimal conditions.

We conclude that spurious and latent formants are the main cause of problems in automatic formant estimation. These problems are quite prevalent, in particular, if the speech samples have been acquired in nonoptimal conditions.

Objectivity of formant determination

As argued above, formants cannot be determined without relying on a priori information about the system (including the "speaker" and the "language") that produced the sounds. In the absence of any supporting information automatic formant analysis is purely objective but the formant frequencies may substantially differ from the values measured by an experienced phonetician. This can, however, be desirable in a setting where the parametrization of the speech signal is more important than a robust correspondence between resonant and formant frequencies (e.g. parametric speech synthesis or speech encoding for mobile phones). However, including *a priori* information about the vocal tract resonances of the vowel articulations or the typical formant values of the vowels in the formant extraction may seem to reduce the objectivity of the analysis.

Formant frequencies and articulation

Across speakers and phonemic contexts, each vowel has a characteristic range of formant frequencies $F_1, F_2, \ldots$ although these ranges overlap as shown in Peterson and

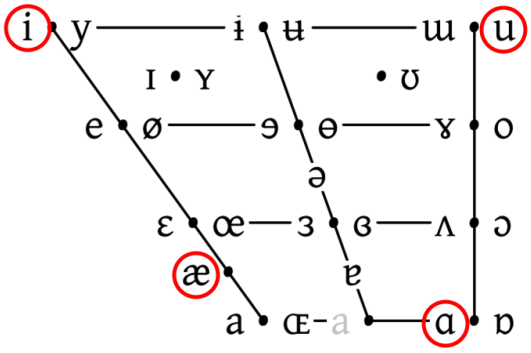

Figure 5. Vowel quadrilateral with all IPA vowel symbols showing the relationship between separate vowel qualities with the articulatory setting of the tongue in terms of height and frontness. The four Finnish vowels depicted in Figures 1 and 4 are circled.

Barney (1952). Typically, the first formant F_1 ranges between 200 Hz and 1200 Hz whereas the second formant F_2 ranges from 600 Hz to 3500 Hz; see (Peterson & Barney, 1952, Figure 8). Within one speaker or one vowel category, the ranges for F_1 and F_2 are much smaller; see Broad (1976). Since formants characterise vowels, the formant frequencies of a vowel sound are expected to be seen at their usual positions if the vowel is known or can be identified by other means. Hence, if a speaker produces a high front vowel, F_1 is expected to be low. If it cannot be observed in the power spectral density (e.g., if f_o is high), the lowest observed formant is nevertheless called F_2 in which case F_1 is regarded latent as explained above.

Vowels can be described through articulatory features such as (i) *tongue height*, (ii) *tongue frontness*, and (iii) *roundedness of lips*. Vowels of a given language can be arranged into distinct clusters in the plane whose coordinate axes correspond to F_1 and F_2, producing the *vowel quadrilateral* shown in Figure 5. The positions of vowels in the quadrilateral bear resemblance to the articulatory organization of vowels in terms of tongue position. In fact, F_1 can be correlated with the tongue height and F_2 with the tongue frontness. The third formant F_3, not shown in Figure 5, is often associated with the roundedness of lips. However, the effect of articulation on formants in full detail is much more complicated as discussed in, e.g., Chiba and Kajiyama (1958), Fant (1958), and Lindblom and Sundberg (1971). When studying speech in clinical populations

having nontypical anatomies or unusual ways of articulation, a careful analysis based on the resonances of the vocal tract air column is thus recommendable.

For F_1, F_2, F_3 and their corresponding Helmholtz resonances $f_{R_1}, f_{R_2}, f_{R_3}$, the modal patterns of the sound pressure vary dominantly in the longitudinal direction of the VT. This is due to the fact that the VT anatomy does not allow standing wave modes in perpendicular directions, i.e., *cross modes*, under frequencies of, say, 4 kHz. For formants at higher frequencies, the relationship between the observed formant frequencies and the computationally obtained VT resonances becomes increasingly complicated and difficult to verbally explain in general terms. For example, the piriform sinuses create antiresonances between 4 kHz and 5 kHz (see Dang and Honda (1997)) as a result of cross modes that are not directly affected by articulation, and, hence, the resulting formant structures in speech are relatively (but not entirely) independent of the vowel uttered. Such phenomena are most conveniently understood by computational 3D acoustic modelling, of which the Helmholtz resonances discussed earlier is a special case. In contrast to the impact of piriform sinuses, Švancara and Horáček (2006) have shown that the impact of tonsilla (or tonsillectomy) to the formant frequencies is vowel specific.

Formants in continuous speech

Normal speech is characterised by continuous movement, where the VT changes shape several times every second. The average peak velocities of the articulators are typically around 0.15 m/s. Thus the resonating air column's geometric shape is rarely static, which in turn is the defining criterion for a resonator. However, compared to the speed of sound – ca. 350 m/s – the articulatory movements are, at least, two magnitudes slower. During a few glottal cycles, the movements are slow enough for VT resonances to occur in a similar manner as in a completely stationary VT. Consequently, and aside from encoding the VT shape, the articulatory movements get encoded into *formant trajectories* in an efficient fashion to the extent that the articulating VT configuration can, at least partially, be reconstructed from known formant frequencies. The relatively

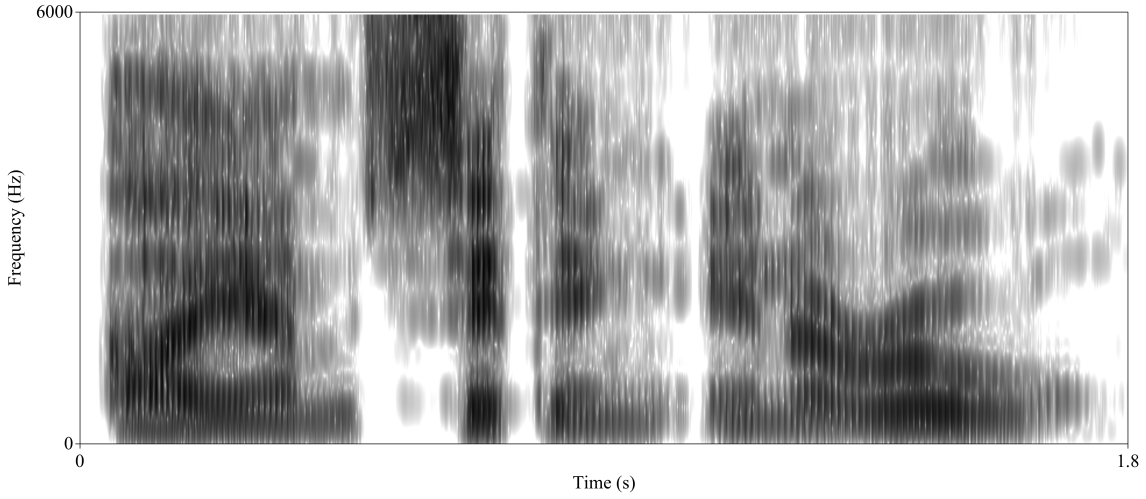

Figure 6. A spectrogram of an English utterance "I am sitting in a room" spoken by a male speaker. The figure clearly shows the dynamic nature of speech as the formants rarely remain stationary.

slow time scales in articulatory movements are reflected in auditory perception where one full period of vowel phonation (of length $T_o = 1/f_o$) may be enough for successful vowel identification; see Robinson and Patterson (1995).

The relationship between articulatory movements and formant frequency values cannot be described in terms of linear response functions. In particular, the acoustic changes get amplified when the articulators are approaching a closure of the VT; see Stevens (1989). The constrictions and closures of the VT during the articulation of most consonants produce so-called *transitions* in the formant trajectories that serve as strong context-dependent cues to the anatomic place of articulation in consonant-vowel sequences as shown in Kewley-Port (1982). The relevance of the formant trajectories can be demonstrated by *sine speech* (see Remez, Rubin, Pisoni, Carrell, et al. (1981)) since the reduction of speech to mere formant structures still maintains some intelligibility.

Further reading

- Chiba and Kajiyama (1958) A foundational work on the mechanism of vowel production and perception from the viewpoints of physiology, physics and psychology.

- Fant (1958) and Fant (1971): Seminal work that used circuits and acoustics to formulate the *source-filter theory of speech production.*
- Stevens (1989): Describes the relationship of articulatory shapes with formants.
- Maurer (2016): A comprehensive work on the acoustics of vowels. Freely downloadable e-book (see link in References).
- Titze et al. (2015): Describes an attempt towards a consensus for symbolic notation for harmonics, resonances, and formants. The system is followed and extended slightly in this article.
- D. Aalto et al. (2014): Description of efficient speech data acquisition during Magnetic Resonance Imaging (MRI), followed by post-processing the speech signals and 3D MR images for formant and Helmholtz resonance analysis.
- Kuortti, Malinen, and Ojalammi (2017): FEM-based Helmholtz resonance analysis of VT vowel geometries is carried out to separate the true and spurious formants from speech spectrogrammes that are recorded during the MRI within the constrained space of the scanner.

Acknowledgments The authors wish to thank M.Sc. J. Kuortti and M.Sc. A. Ojalammi for producing the data and the illustrations, and Dept. Signal Processing and Acoustics at Aalto University (Prof. P. Alku), PUMA research group at Dept. Oral and Maxillofacial Surgery, University of Turku (Prof. R.-P. Happonen), and Medical Imaging Centre of Southwest Finland (Prof. R. Parkkola and Dr. J. Saunavaara) for cooperation in MRI and speech data acquisition.

References

Aalto, A., Lukkari, T., & Malinen, J. (2015). Acoustic wave guides as infinite-dimensional dynamical systems. *ESAIM: Control, Optimisation and Calculus of Variations*, *21*(2), 324–347. (Published online: 17 October 2014) doi: http://dx.doi.org/10.1051/cocv/2014019

Aalto, D., Aaltonen, O., Happonen, R.-P., Jääsaari, P., Kivelä, A., Kuortti, J., ... Vainio, M. (2014). Large scale data acquisition of simultaneous MRI and speech. *Applied Acoustics*, *83*(1), 64–75.

Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M., & Story, B. (2013). Formant frequency estimation of high-pitched vowels using weighted linear prediction. *Journal of the Acoustical Society of America*, *134*(2), 1295-1313.

Arnela, M., Guasch, O., & Alías, F. (2013). Effects of head geometry simplifications on acoustic radiation of vowel sounds based on time-domain Finite Element simulations. *Journal of the Acoustical Society of America*, *135*(4), 2946–2954.

Broad, D. (1976). Toward defining acoustic phonetic equivalence for vowels. *Phonetica*, *33*(6), 401–424.

Burg, J. (1975). *Maximum entropy spectral analysis.* Unpublished doctoral dissertation, Stanford University.

Chiba, T., & Kajiyama, M. (1958). The vowel: its nature and structure (1942). *Setagaya: Phonetic Society of Japan, Tokyo.*

Dang, J., & Honda, K. (1997). Acoustic characteristics of the piriform fossa in models and humans. *The Journal of the Acoustical Society of America*, *101*(1), 456–465.

El-Jaroudi, A., & Makhoul, J. (1991). Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, *39*, 411âĂŞ-423.

Epps, J., Smith, J., & Wolfe, J. (1997). A novel instrument to measure acoustic resonances of the vocal tract during phonation. *Measurement Science and Technology*, *8*(10), 1112.

Fant, G. (1958). *Acoustic theory of speech production.*

Fant, G. (1971). *Acoustic theory of speech production: with calculations based on x-ray*

studies of russian articulations (Vol. 2). Walter de Gruyter.

Fulop, S. (2011). *Speech spectrum analysis*. Springer Science & Business Media.

Hannukainen, A., Lukkari, T., Malinen, J., & Palo, P. (2007). Vowel formants from the wave equation. *Journal of the Acoustical Society of America Express Letters*, *122*(1), EL1–EL7.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant–vowel syllables. *Journal of the Acoustical Society of America*, *72*(2), 379–389.

Kuortti, J., Malinen, J., & Ojalammi, A. (2017). Post-processing speech recordings during MRI. *Biomedical Signal Processing and Control*. (to appear)

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

MacDonald, E. N., Purcell, D. W., & Munhall, K. G. (2011). Probing the independence of formant control using altered auditory feedback. *Journal of the Acoustical Society of America*, *129*(2), 955–965.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of IEEE*, *63*, 561–580.

Maurer, D. (2016). *Acoustics of the vowel*. Bern, Switzerland: Peter Lang. Retrieved from `//www.peterlang.com/product/47242`

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

Remez, R. E., Rubin, P. E., Pisoni, D. B., Carrell, T. D., et al. (1981). Speech perception without traditional speech cues. *Science*, *212*(4497), 947–949.

Robinson, K., & Patterson, R. D. (1995). The stimulus duration required to identify vowels, their octave, and their pitch chroma. *Journal of the Acoustical Society of America*, *98*(4), 1858–1865.

Stevens, K. N. (1989). On the quanta! nature of speech. *Journal of phonetics*, *17*, 3–45.

Stoica, P., & Moses, R. (2005). *Spectral analysis of signals*. Prentice Hall.

Švancara, P., & Horáček, J. (2006). Numerical modelling of effect of tonsillectomy on production of czech vowels. *Acta Acustica united with Acustica*, *92*(5), 681–688.

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