SEGMENTAL ACOUSTIC PHONETICS I

1. **Aims**

1. Within a phonetic framework, to introduce the acoustic properties of speech sounds in isolation or in simple syllabic contexts, prior to considering them in connected speech.
2. To draw preliminary connections between the acoustic-phonetic properties of speech and speech production and perception.

2. **Purpose of study**

*Agreement:* You need to know what’s in the acoustic signal if you want to understand how it’s produced and/or how it’s understood.

*Questions:* What are the important units—words? morphemes? syllables? phonemes? phones? features? diphones ??

*Must grammatical and prosodic structure always be acknowledged in phonetics?*  
*Is there more than one unit of production/perception?*  
*Are any (or all) units invariant across contexts, speakers?*

*cf. that man vs the man that I saw: */ðæt/ vs. */ðæt/*  
*he diced them vs. he’d iced them: shorter and ‘weaker’ */d/* syllable-finally than syllable-initially  
*I sing descant vs. I was singing descant: how much can the final */n/* assimilate to the */d/*?

3. **Review: three acoustic representations of sound**

<table>
<thead>
<tr>
<th></th>
<th>horizontal axis</th>
<th>vertical axis</th>
<th>blackness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spectrogram:</strong></td>
<td>time</td>
<td>x</td>
<td>frequency</td>
</tr>
<tr>
<td><strong>Waveform:</strong></td>
<td>time</td>
<td>x</td>
<td>amplitude</td>
</tr>
<tr>
<td><strong>Spectrum:</strong></td>
<td>frequency</td>
<td>x</td>
<td>amplitude</td>
</tr>
</tbody>
</table>

4. **Review: Types of speech noise, as visible in wideband spectrograms**

Broad phonetic **MANNER** of articulation (e.g. stop / fricative / nasal) is usually relatively clearly evident in acoustic records of speech, and is related to **excitation type** (periodicity (phonation) / turbulence noise / transient excitation / mixtures of these).  
*Cues to phonetic **PLACE** of articulation are often less easily captured.*

*Figure 1.* Upper panel shows file ~paper9/demo/cups.d  (Lower panel is ~paper9/demo/away.d)

**Review of representation of excitation type:**

*Phonation:* vertical black striations at more or less regular time intervals: each striation reflects the noise made during a single glottal pulse.

*Fricative noise:* aperiodic noise showing no regularities in the time domain; it varies in darkness *(amplitude)* from very black *(high amplitude)* to faint grey *(low amplitudes)*; it may be concentrated in a few frequencies, or spread fairly evenly across all the frequencies. In speech, if it’s concentrated in only some frequencies, they are usually—but not always—the high ones.

*Stop bursts:* sharp line *(transient)*, sometimes like a pencil-line *(especially for bilabials).* This represents the abrupt release of high-pressure air as the articulators rapidly open at the end of the closure of an oral stop. The shape of the transient in the frequency domain depends on the place of articulation of the stop.

*Silence:* white space: no excitation, i.e. no source of sound.

*Mixed excitation:* a mixture of phonation with fricative noise or stop burst e.g. in [v].
In fluent connected speech (i.e. excluding silent pauses between phrases, and silent hesitations, etc):
- There can be many silences, but they are associated with the closure periods of voiceless stops.
- There are no silences between words unless the words begin and/or end with voiceless stops.
- There is no 1:1 correspondence between phoneme-sized ‘phonetic segments’ and sounds:
  - some phoneme-sized segments are obvious in acoustic representations, some are not;
  - some acoustic segments correspond closely to phones, some do not.

Syllable centers are obvious when they are bounded by obstruents (stops or fricatives), but more difficult to identify unambiguously when they contain only approximants and vowels:

\[
\begin{array}{c}
\sigma \\
\text{Onset} \\
\text{Rhyme} \\
| \text{Nucleus} | \text{Coda} \\
\end{array}
\quad
\begin{array}{c}
\sigma \\
\text{Onset} \\
\text{Rhyme} \\
| \text{Nucleus} | \text{Coda} \\
\end{array}
\]

\[
\sigma \quad \sigma \\
\text{Onset} \quad \text{Onset} \\
\text{Rhyme} \quad \text{Rhyme} \\
\text{Nucleus} \quad \text{Nucleus} \\
\text{Coda} \quad \text{Coda}
\]

\[
b \quad a \\
\text{Nucleus} \\
\text{Coda}
\]

\[
w \quad 3 \\
\text{Nucleus} \\
\text{Coda}
\]

cf. Figure 1 upper vs lower panels. Listen to them on SGs: ~paper9/demo/cups.d and away.d

5. Isolated vowels use ~paper9/demo/vowels/hdmonophths.fn.d (or iauai.fn.d or hdvowels.fn.d)

As discussed in earlier sessions, for a given speaker, the shape of the vocal tract crucially determines the sound quality by affecting the pattern of resonance frequencies, or formants. The sound quality changes because the vocal tract’s resonance (formant) frequencies shift when it is constricted. Each different constriction location in the vocal tract produces its own distinctive vowel sound, because it produces its own distinctive resonance pattern.

Fig. 2.

remember this sketch:

\[
\begin{array}{c|c}
\text{Hz} & \\
\hline & F2 \\
\hline & F1 \\
\end{array}
\]

\[
\text{Vowel quality heard when the relative frequencies of F1 and F2 are those immediately above each symbol}
\]

Use this sketch like a mnemonic to remember relative frequencies of the first two formants for different vowels.

Then learn some representative frequencies for /i a u/, from which you can roughly estimate the quality of other vowels. This will be useful for all lab work and your understanding of theory.

Interpreting spectrograms of vowels in terms of phonetic knowledge of tongue, jaw and lip position allows us to roughly see how formant frequencies reflect vocal-tract shape.

Fig. 3: some English vowels. Acoustically:
- tongue lowering → high first formant (F1);
- tongue fronting → high second formant (F2).
- jaw opening → greater overall intensity (and high F1).
- lip rounding → lowered formants, especially (usually) F2.

Fig. 4:
6. **Isolated voiceless fricatives**

*Figure 5: Some voiceless fricatives*

Voiceless fricatives are produced without phonation—no vocal fold vibration—entirely aperiodic. The sound source arises in the vicinity of a narrow constriction formed by the articulators; the constriction causes the air to rush through it, which creates turbulence noise. The turbulent airstream may also hit another obstruction; for example, the teeth are necessary to make a sibilant [s] sound.

When the major constriction is inside the mouth, the resultant sound is relatively loud, with mainly high-frequency energy; i.e. the top half of the spectrogram is relatively black. In acoustic terms, we say these voiceless fricatives have intense, or high amplitude, high-frequency energy.

For these fricatives, the further the major constriction is from the lips, the lower the frication noise extends in frequency: compare [s] vs [ç] vs [x], with major constrictions successively further back in the mouth. This is because most of the excitation is due to the length of the cavity in front of the major constriction. Note clear formants for the more back fricatives. English [ʃ] is like [s] in having high-amplitude aperiodicity in the high frequencies, but [ʃ] is distinctive in that the greatest amplitude of aperiodicity is in the 2-4 kHz region. There are a number of reasons for this, one being that English [ʃ] is lip-rounded; rounding/protruding the lips lowers spectral frequencies, especially in the mid-frequencies.

Fricatives made at the lips have much lower-amplitude (quieter) energy, which is usually more evenly distributed across the entire frequency range, because there is little or no tube in front of the labial constriction. Compare [f] with the others. However, when the lips are protruded, labial fricatives may show a formant structure—i.e. uneven distribution of energy across the frequency range.

Voiced fricatives have mixed periodic and aperiodic excitation because the vocal folds vibrate while turbulence noise is also generated at a constriction elsewhere in oral cavity. To maintain this type of takes quite a bit control of subglottal air pressure and vocal fold configuration. The balance of between the aperiodic and periodic excitation varies between languages, styles of speech, and phonetic contexts. English voiced fricatives often have relatively light (even absent) periodicity. There are other cues though. E.g. voiced fricatives are usually shorter than voiceless ones, and a slower rate of decay of the amplitude envelope, or low frequencies, into the obstructed closure. See [v] vs [f] in *cups.d*.
Isolated(?) oral stops (oral stops in simple vocalic contexts)

An intervocalic stop can have up to 5 distinct articulatory-acoustic phases:

<table>
<thead>
<tr>
<th>articulation</th>
<th>acoustic consequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 vocal tract closing</td>
<td>transitions into the closure</td>
</tr>
<tr>
<td>2 closure</td>
<td>silence or near-silence</td>
</tr>
<tr>
<td>3 release</td>
<td>burst, or transient + frication noise</td>
</tr>
<tr>
<td>4 aspiration</td>
<td>voiceless vowel sound</td>
</tr>
<tr>
<td>5 vocal tract opening</td>
<td>transitions out of the closure</td>
</tr>
</tbody>
</table>

Which phases you actually get depends on the position of the stop in the broader linguistic structure, and the way that the speaker actually pronounces it, which can vary e.g. with degree of casualness. Here, we consider only clearly-pronounced stops in syllable onsets, in CV and VCV contexts. Notice that stages 1, 4, & 5 are as much part of the vowel as of the stop.

Fig. 6: Spectrograms of some voiceless and voiced stops—necessarily in a context, here [Ca], thus not really “isolated segments”.


Fig. 7: Schematic spectrograms of intervocalic /b d g/, with the vowel /a/, showing typical transition patterns in the first three formants, and approximate transient patterns of the transients (up to about 4 kHz). Stages 1, 2, 3 & 5 are present. These transition patterns are reliably found in the [a] context and often other contexts too. They are worth memorizing.
But Figures 8a and b illustrate that other vowel contexts don’t necessarily produce the same transition patterns. The relationship between the burst transients and the following formants is somewhat more constant than the formant transition patterns in the vowels. These changes are typical examples of the way phonetic context affects the acoustic pattern of a ‘canonical’ (textbook) utterance.

(i) Schematic spectrograms of VdV utterances.  
(ii) x-ray of vocal tract at middle of /d/ occlusion. 
(iii) x-ray of vocal tract in vowel steady state.

These pictures show different formant patterns due to different vocal tract shapes during articulation of the [d], which in turn are due to coarticulatory differences due to vowel context.


Fig. 8b. Spectrograms of voiced stops in CV syllables with vowels /i a u/, illustrating differences in formant transitions

[MML L9A.M.3]
The transition pattern is an important perceptual cue for place of articulation, but the relationship is not simple, as the phenomenon of **categorical perception** illustrates:

When small changes are made in the frequencies of F1 or F2 of a vowel, we always tend to hear a shift in vowel quality, as long as the change is big enough to be perceivable. But for consonants, it is possible to make changes that are all of equal sizes, and yet we hear some of those changes very clearly, and others we may not notice at all. This uneven perception of equal acoustic changes is known as categorical perception. It is illustrated schematically in Figure 9.

**Figure 9.**

Equal acoustic changes

```
```

Percept

Categorical perception was originally demonstrated *only* for (synthetic) stop consonants:
- voicing of stops: /b/ vs /p/; /d/ vs /t/; /g/ vs /k/.
- place of articulation of stops: /p/ vs /t/ vs /k/; /b/ vs /d/ vs /g/.

**Figure 10.** Schematic synthetic stimuli used to demonstrate categorical perception for **place of stop articulation**. 14 two-formant stimuli were drawn on the Haskins Pattern Playback. F1 was the same for all stimuli. The F2 transition was the only thing varied. From Liberman et al. (1957)

**Fig. 11.** Identification and two-step discrimination functions for one listener.

*Left panel:* percentage of time each stimulus was identified as /b/, /d/ or /g/. *Right panel:* result of the two-step discrimination test, using the technique described in the text of the article. (Adapted from Liberman et al.: J. Experimental Psychology, 54 © 1957, American Psychological Assoc.

Experiments demonstrating categorical perception usually show steep identification functions, and, in discrimination functions, peaks between stimuli that span the (50%) identification boundary.

N.B. The focus so far has been on stop properties due to changing vocal tract shape. Other aspects of stops, including laryngeal contributions to stops, and perceptual cues, will be addressed later.
7. Approximants in simple vocalic contexts

Approximants (liquids e.g. [l] [r], and glides e.g. [j] [w]) have a more extreme constriction than vowels, but not so extreme as to create turbulence. The definition of approximants is partly based on phonological criteria: they occur at syllable margins, not as syllable nuclei. Nonetheless, their acoustic correlates are distinctive enough, provided you interpret them relative to the surrounding context:

- lower overall energy than in vowels (because the vocal tract is less open)
- more rapid formant movements than vowels, but less rapid than in obstruents (stops/fricatives)

Figure 12 shows x-ray tracings of American English approximants. Figure 13 shows formant patterns for syllable-initial approximants in two vocalic contexts.

1. Note the similarity of articulation of [j] and [w] to [i] and [u] respectively, resulting in similar formant patterns (Figure 13): in [j], F2 is high and close to F3; in [w], F2 is low and close to F1.
2. The lateral approximant [l] has a central, tongue-tip constriction at the alveolar ridge. Air flows around the sides of the tongue; this airflow pattern gives rise to complex acoustics. At the release of the constriction, a transient (like a stop burst) is sometimes visible in the spectrogram (see Lee in Figure 13). These are syllable-onset [l]s; expect different patterns for syllable-coda [l].
3. [r] is unusual in that it lowers F3, F4, F5 considerably. Other rhotic sounds (e.g. labiodental or uvular) achieve similar acoustic effects by different means.
Rate of articulator movement, and its acoustic correlate the duration of transitions, is critical to whether a sonorant sounds of this type are heard as approximants or vowels. More generally, many phonemic distinctions rest upon relative differences in duration, e.g., stop vs approximant vs vowel. The human perceptual system seems to make decisions based on relative rate (Miller 1981 *Phonetica*); presumably computations are continuously updated, since rate can vary in one utterance. In the experiment using the stimuli illustrated in the next Figure (12), lengthening the duration of the same frequency transition in a CV syllable changed the percept from stop to approximant.

![Schematic spectrograms of stimuli differing only in rate of initial syllable transition.](image)

**Figure 14**: Schematic spectrograms of stimuli differing only in rate of initial syllable transition. That is, the frequency range of the transition is identical in all stimuli, but the timing differs between stimuli. Steady state portions of the vowels are incompletely shown, except for [ue] and [ie]. In the experiment using the stimuli illustrated here, lengthening the duration of the same frequency transition in a CV syllable changed the percept from stop to approximant. From Liberman et al. (1956), *J. Exp. Psych.* 52.

### 8. Summary

1. Approximate syllable structure, gross prosodic properties, and broad phonetic categories (MANNER of articulation) are usually relatively clearly evident in waveforms and spectrograms.
2. Cues to place of articulation mainly appear as particular distinctive relationships in formant frequency over time, i.e., in formant frequency pattern, but these cues to place are often harder to identify with certainty than those for manner of articulation.
3. The ear/brain’s interpretation of the changes in the acoustic speech signal is sometimes relatively straightforwardly related to those changes, and sometimes remarkably complex.
4. Thus one of the greatest challenges is to work out what ‘units’ of perception or production we should be focussing on. This is especially interesting if you want a theory or model that can account for a wide range of normal styles of speech. This will be discussed in later lectures. If you want to start thinking about it now, here are some pointers, which you can use to come to your own conclusions or discuss in supervisions:

- **Figure 15**: isolated “phoneme-sized segments” i.e. phones vs connected speech for the phones of *steal*.
• **Figure 16**: moderately slow, formal pronunciation and fast, informal pronunciation of *did you eat?*

9. **Reading**

**General sources**: minimally, read one of the following, (or Ladefoged and one other):


**Older but still useful** (speech hasn’t changed that much):


**More advanced**:


- Malmberg, B. (Ed.) (1968). *Manual of Phonetics*. Amsterdam: North Holland. Ch 8 (Fant), pp 236-253 only. Excellent spectrographic material. (The rest of the chapter is on the acoustic theory of speech production — excellent, but very technical.)
