EECS225d Audio Signal Processing in Humans and Machines **Spring 1995**

Lecture 22: March 10, 1995

Lecturer: Jont Allen *Scribe: Nikki Mirghafori*

22.1 How Do Humans Process and Recognize Speech?

Jont Allen's talk was based on his recent paper in IEEE Transactions on Speech and Audio Processing [A94] (handed out in class). In this paper, Allen reviews the work of Harvey Fletcher, who started the articulation index (AI) proprietary project at AT&T Bell Labs in about 1918.

22.1.1 Motivation

The motivation was to quantify the quality of speech sounds in the telephone plant to improve speech intelligibility and preference. Today, this study is particularly relevant to the relatively new science of automatic speech recognition.

22.1.2 Highlights

- We recognize speech based on a cascade of layers
- Context effects are strong and confound the study of recognition
- Removing context factors from the human speech recognition problem, increases testing efficiency, and reduces test variability
- The phone (one allophone of a phoneme) seems to be the basic unit of speech
- The phone is derived from independent auditory and visual channels
- Each channel depends on the SNR, not the energy
- To account for independent articulation channels, we must be extracting local features which are integrated in the brain
- Using context free databases in ASR should be fruitful

22.1.3 The Model

The proposed model for human speech processing (see Fig. 6 in [A94]) has five layers, all of which are below the language level. There is no feedback in this simplified model. The first layer is the cochlear filter bank, which affects the SNR. The next level is the *feature extraction* layer, where the input is SNR_k and output is a partial-phone recognition, $1 - e_k$. Third layer maps the features to phones. In the last two layers, syllables and words are found.

22.1.4 Controlling Context Entropy

The context entropy of a CVC (constant-vowel-constant, e.g., *fit*) word is lower than that of a nonsense CVC (e.g. *fif*); in other words, context is a strong cue for speech understanding. Fletcher controlled this variable in the articulation index experiments by using nonsense CVC's. If we use the phone frequencies given by Fletcher, the average phone entropy is 4.3 bits/phone. This entropy would drop as context effects are included.

Note that Fletcher used the word *articulation* to mean the probability of correctly identifying nonsense speech sounds, while *intelligibility* is the probability of correctly identifying meaningful speech sounds, such as words.

22.1.5 The Articulation Experiment

Fletcher used balanced nonsense syllables (CV, VC, & CVC) to minimize context effects. He held the speech corpus constant for each experiment in order to maintain constant source entropy. He computed an average over many caller-listener pairs. The articulation was varied by: 1) changing the SNR, and 2) Low Pass & High Pass filtering the speech.

The syllable lists were spoken and the listeners recorded what they heard (see Fig 1 in [A94]). Then, error probabilities $1-c$ and $1-v$ for the consonant and vowel sound-units, respectively, were computed. To simplify computations, an average {C,V} speech-unit articulation probability s was computed from the composition of {C,V} units in the database (i.e., $s = \frac{(2c+v)}{3}$ for CVC's). For perfect conditions (i.e., high SNR, no HP/LP filtering) the recognition accuracy was 98.5%.

22.1.6 The Results of the Experiment

Fletcher empirically found that the CVC syllable articulation $S(\alpha)$, where α is the gain applied to the speech, is accurately predicted from the phone articulations $c(\alpha)$ and $v(\alpha)$ by the relation $S(\alpha) = c(\alpha)^2 v(\alpha) \approx s(\alpha)^3$; this implies that the three sound-units are heard as independent sounds. In other words, the probability of correct recognition for the average *phone* determines human nonsense syllable recognition. Note however that this is only true for nonsense words, since for meaningful words context will decrease the entropy.

After having found the fundamental importance of the phone articulation, s, to human speech recognition, Fletcher further studied s for various channel frequency responses and channel noise, using LP and HP filters on the speech. He found that the *partial articulations*, s_L and s_H, articulations for low and high band, respectively, did not sum to the wide band articulation s. Hence, he looked for a non-linear transformation of the partial articulations that would make them additive, as in

$$A(s_L(f_c,\alpha)) + A(s_H(f_c,\alpha)) = A(s(\alpha)) \tag{22.1}$$

where f_c is the high- and low-pass cut-off frequency. Fortunately Fletcher found $A(s)$, which he called *articulation index*, empirically:

$$A(s) = \frac{log_{10}(1-s)}{log_{10}(1-s_{max})} \tag{22.2}$$

where constant s_{max} is the maximum articulation and $e_{min} = 1-s_{max} = 0.015$ is the corresponding minimum articulation error. Solving equation 22.2 for s, we get $s(A) = 1 - e_{min}^A$ which could be written in terms of $e(A)$ simply as $e(A) = e_{min}^A$.

The nonlinearly transformed articulation defines an *articulation index density* $D(f)$ over frequency f. Integration or summation over this density gives the *articulation index* A. A can be viewed as a fundamental internal variable of speech recognition. $A(s)$ could be calculated using the articulation index density:

$$A(s_L(f_c)) = \int_0^{f_c} D(f)df \tag{22.3}$$

$$A(s_H(f_c)) = \int_{f_c}^{\infty} D(f)df \tag{22.4}$$

$$A(s) = \int_0^{\infty} D(f)df \tag{22.5}$$

where the density $D(f)$ is uniquely determined from:

$$D(f_c) = \frac{\partial}{\partial f_c} A(s_L(f_c)) \tag{22.6}$$

Fletcher derived the *density over frequency of the phone articulation index* $D(f)$ from his experiments.

22.1.7 Independent Articulation Bands

Then Fletcher showed that phones are processed in independent *articulation bands* (frequency channels), and that these independent estimates are "optimally" merged. For example if we have 100 spoken sounds and 10 errors are made while listening to the low band, and 20 errors are made while listening to the high band, then $e = 0.1 * 0.2 = 0.02$, that is, only 2 errors will be made when listening to the full band and $s = 1 - 0.02 = 0.98$. This result was showed through the following observations: From equations 22.1 and 22.2 we get

$$log(1 - s) = log(1 - s_L) + log(1 - s_H) \tag{22.7}$$

which becomes

$$1 - s = (1 - s_L)(1 - s_H) \tag{22.8}$$

and in terms of the articulation error $e = 1 - s$

$$e = e_L e_H \tag{22.9}$$

This equation is true for any value of f_c. It says that we are listening to *independent* sets of phone features in the two bands and processing them independently, up to the point where they are fused to produce the phone estimates. Fletcher expanded the two band example above to the multi-channel model to as many as $K = 20$ bands, where $1 - s = e_1 e_2 e_3 ... e_k ... e_K$.

Taking $K = 20$ makes each band correspond to 1 mm along the basilar membrane. Since Fletcher identified a critical band to be about 0.5 mm along the basilar membrane, one articulation band represents two critical bands. Figure 5 in [A94] shows the ratio of $D(f)/kf$ where $D(f)$ is the articulation density and $k(f)$ is the critical ratio (a measure of the relative bandwidths of our cochlear filters). This plot of $D(f)/kf$ represents the measure of relative speech articulation per critical band and it is approximately uniform over the speech band.

22.1.8 Band Error e_k and SNR

They found that when the SNR is varied, the articulation band error is given by the relation

$$e_k = (e_{min})^{SNR_k/30K} \tag{22.10}$$

where SNR_k is the signal to noise ratio in dB for each band k. Each band k, where $K = 20$ is the total number of bands, corresponds to 1 mm along the basilar membrane. Also recall from before that e_{min} is 0.015. This relationship tells us that it is the speech SNR *and not the energy* which determines the phone articulation s.

Jont Allen noted a serious error in the model in the limit. For very low SNR_k, we expect that articulation should go to chance: $e_k \longrightarrow 1 - 1/M \Longrightarrow s_k \longrightarrow 1/M$ where M is determined by the number of bits/band. However, according to the model, it goes to zero: $e_k \longrightarrow 1 \Longrightarrow s_k \longrightarrow 0$.

22.1.9 The Recognition Chain

The recognition chain in the model is the following:

- The signal enters the cochlea and is broken into critical bands which define the signal-to-noise ratios SNR_k, where k labels the cochlear frequency channel.
- The first layer processes the output from the above and defines the phone features represented by partial articulation errors e_k, using equation 22.10.
- Then, independent-band phone articulation model $s = e_1 e_2 ... e_k .... e_K$ determines s.
- The phones are transferred into (nonsense) CVC syllable units with articulation $S = s^3$.
- Words are determined with intelligibility $W(A) = 1 - (1 - S(A))^j$, where the constant $j > 1$ depends on the entropy of the word corpus and may be empirically determined.

Allen argues that any feedback mechanism could create serious "real-time" problems, so it is unlikely that feedback is common or significant between the deeper and outer layers.

22.1.10 The Phone Feature Space?

Allen uses the term *feature* to refer to 1 bit of partial recognition, as a binary concept. Using $K = 20$, each band corresponds to 1 mm along the basilar membrane. Given that the distance between 300 Hz and 8 kHz on the basilar membrane is 20 mm and the phone entropy is 4.3 bits (see above), there are $20/4.3 = 4.65$ mm/feature. Furthermore, since 1 octave corresponds to about 5 mm along the basilar membrane, the average phone feature density is about 1 feature/octave and the average feature length is 1 octave/feature.

Allen argues against a template-based (*across-frequency* processing) approach to phone features extraction and espouses a locally-based *across-time* processing: First, cochlea breaks the external world down into a tonotopic array of critical bands. The CNS "reconstructs" the scene from these pieces by making a huge cascade of binary decisions. The local feature extractors define the feature length along the tonotopic axis. The number of correlated feature regions determines the dimension of the space and the coordinates (e.g., the probability of the feature being present). Phone recognition is not the synchronous timing of feature events, but some more abstract relation between the presence and absence of features, and their geometrical relations in a multidimensional feature space.

References

[A94] JONT ALLEN *How Do Humans Process and Recognize Speech?*, IEEE Transactions on Speech and Audio Processing, Oct 1994, pp 576-576.