TIME-FREQUENCY DISTRIBUTION OF PARTIAL PHONETIC INFORMATION MEASURED USING MUTUAL INFORMATION

Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering,
University of Illinois, Urbana, IL 61801, USA

ABSTRACT

This paper uses the method of mutual information to estimate the distribution of partial phonetic information in the time-frequency plane relative to an acoustic landmark. TIMIT transcriptions are parsed to estimate the locations of consonant closure landmarks, consonant release landmarks, manner change landmarks, and vowel or glide pivot landmarks. A mel-scale spectrogram is computed over the 250ms centered at each landmark, and the logarithmic energy of each point in time-frequency space is linearly quantized. The phoneme label associated with a landmark determines the values of 25 binary distinctive features. Finally, coincidences between feature and spectral energy values are counted, and the average log probabilities are calculated in order to produce an "infogram" of each distinctive feature: a measurement of the mutual information between the value of the feature and the energy of each point in the time-frequency plane.

1. INTRODUCTION

Acoustic phonetic experiments indicate that the information content of speech is concentrated around salient acoustic landmarks: the release and closure of a consonant, the maximal constriction of a glide, and the most steady-state portion of a vowel [5]. Stevens et al. proposed the knowledge-based alignment of distinctive features with acoustic landmarks [7], and Liu was able to detect acoustic landmarks using a knowledge-based system with an accuracy of about 91% [3]. Halberstadt found that a statistical landmark-based recognition architecture yields higher word accuracy than a segment-based architecture [2, 1].

This paper supplements the work of Stevens et al. by using the mutual information methods of Yang, van Vuuren, and Hermansky [9], applied to a database of 1491 TIMIT sentences. In this paper, mutual information is computed between a distinctive feature specified at a landmark and each spectral amplitude $X(t, f)$ in a mel-frequency spectrogram centered at the landmark. The resulting mutual information measure, which this paper calls an "infogram," provides a preliminary information-theoretic estimate of the distribution of spectral information about distinctive features in the time-frequency plane relative to a landmark.

2. METHODS

The infogram is an estimate of the mutual information between the value of a distinctive feature and the amplitude of each point in time-frequency space. In order to compute mutual information, it is necessary to know and distinctive feature values of several thousand landmarks in a database of continuous speech. Section 2.1. describes methods for estimating the locations of landmarks and the values of distinctive features based on the segmental transcriptions provided in the TIMIT database.

Landmark	Phone Class	Context	Position
Closure	S,F,N	V,G left	Start
Manner Change	S,F,N	S,F,N right, diff. manner	End
Release	S Release	V,G right	Start
	N	V,G right	End
	F	V,G right	End-20%
Pivot	V	all	Start+33%
	G	Intersyllabic	Start+50%
	G	Syl. onset	Start
	G	Syl. coda	End

Table 1. Heuristic rules for estimating landmark locations based on TIMIT transcriptions. S=stop, F=fricative, N=nasal, V=vowel, G=glide, liquid, flap, or /h/.

2.1. Distinctive Feature Transcription

This paper makes use of four types of acoustic landmarks: consonant closure landmarks, consonant release landmarks, vowel and glide pivot landmarks, and manner change landmarks. A closure landmark consists of the closure from a vowel, glide, or liquid into an obstruent or nasal consonant. Manner-change landmarks consist of transitions between two consonants which differ in the features sonorant, continuant, or strident. Glide and flap pivot landmarks are an estimate of the point of maximum constriction of a glide, liquid, or flap consonant, and vowel pivot landmarks are an estimate of the perceptual center of the vowel (often but not always the point of maximum opening).

The locations of closure, release, and manner change landmarks are transcribed in the TIMIT database [10]. The locations of pivot landmarks are not transcribed, and must be estimated on the basis of the transcribed segment start and end times. In this work, the locations of all landmarks in the TIMIT TRAIN database were estimated by parsing segmental transcriptions using the heuristic rule set shown in table 1. A TIMIT segment of the specified type, in the specified context, generates a landmark of the type given in the first column. The position of the landmark relative to the transcribed start and end times of the segment is also specified.

Distinctive features were assigned to each phoneme on the basis of a nonlinear distinctive feature geometry similar to that proposed by Stevens et al. [7]. Theories of nonlinear phonology define a hierarchy of binary distinctive features, in which each distinctive feature d may be either present ($d = +1$), absent ($d = -1$), or unspecified. A distinctive feature is unspecified only if the settings of its parent features determine that the feature is either meaningless or not linguistically salient. In the feature hierarchy used here, the root feature [consonantal] is always specified. All [+consonantal] segments have a specified manner (consisting of the features [sonorant] and [continuant]) and place of articulation (consisting of the features [lips], [blade], and [body]); other features may be specified or unspecified, depending on the manner and place of articulation. In addition to the minimum set of features suggested by Stevens et al., a number of redundant distinctive features were investigated, in order to test the division of phonemes into non-standard subsets. For example, the features [grave] and [compact] label place of articulation based on acoustic rather than articulatory criteria.

2.2. Calculation of Mutual Information

A 51-frame, 30-band, 23-level quantized mel-scale spectrogram was computed centered at each estimated landmark. Three analysis window lengths were tested (6ms, 12ms, and 20ms), and windows were shifted 5ms between spectral slices. Each spectral slice was computed using a 1024-point FFT estimate of the power spectrum. Power spectral samples were combined using a bank of Hanning-shaped filters, with center frequencies and bandwidths uniformly spaced on a mel-frequency scale between 0 and 2840 mel (8kHz). The TIMIT TRAIN database was searched for the minimum and maximum values of each of the 30 log-compressed spectral samples, and a linear quantization scheme was designed by dividing the range of log-amplitudes of each spectral sample into 23 equal increments.

Coincidences between feature and spectral energy values were counted across 1491 sentences (one "si" and two "sx" sentences from each speaker in TIMIT TRAIN). Specifically, the number $N_{tfLD}(x,d)$ of co-occurrences of spectral level $x = X(t,f)$ and feature value $d \in \{-1,1\}$ was counted separately for each time t, frequency f, landmark type L, and distinctive feature type D. The probability of co-occurrence was estimated as

$$p_{tfLD}(x,d) = \frac{N_{tfLD}(x,d)}{\sum_{x=1}^{23}\sum_{d=-1,1} N_{tfLD}(x,d)} \qquad (1)$$

The mutual information between the distinctive feature D and each point in time-frequency space was computed by adding log-probabilities, resulting in the infogram $I_{LD}(t,f)$:

$$I_{LD}(t,f) = \sum_x \sum_d p(x,d) \log_2 \left(\frac{p(x,d)}{p(x)p(d)} \right) \qquad (2)$$

where subscripts on the probabilities have been dropped for convenience.

$I_{LD}(t,f)$ is the difference between the *a priori* entropy of the distinctive feature, H_{LD}, and its conditional entropy given knowledge of the spectral energy $X(t,f)$:

$$I_{LD}(t,f) = H_{LD} - H_{LD|X}(t,f) \qquad (3)$$

where

$$H_{LD} = -\sum_d p(d) \log_2 p(d) \qquad (4)$$

$$H_{LD|X}(t,f) = -\sum_x p(x) \sum_d p(d|x) \log_2 p(d|x) \qquad (5)$$

H_{LD} and $H_{LD|X}(t,f)$ are both non-negative, and $H_{LD|X}(t,f) \leq H_{LD}$. A specified distinctive feature has only two possible values ($d = 1$ and $d = -1$), so, from the equations above, it follows that

$$0 \leq I_{LD}(t,f) \leq H_{LD} \leq 1 \qquad (6)$$

$I_{LD}(t,f) = 0$ if the spectrum is independent of the distinctive feature, and $I_{LD}(t,f) = H_{LD}$ if the spectrum completely determines the distinctive feature. $H_{LD} = 1$ if the two values of the distinctive feature ($d = -1$ and $d = 1$) are equally likely *a priori*.

3. RESULTS

The maximum amount of information about a distinctive feature which is provided by any one spectral sample is

$$I^*_{LD} = \max_t \max_f I_{LD}(t,f) \qquad (7)$$

Table 2 lists the average value of I^*_{LD} for each feature D, computed using a 20ms spectral analysis window, and averaged over all landmark types at which D is distinctive. I^*_{LD} is measured both in bits, and as a fraction of the *a priori* feature entropy H_{LD}. The global average of I^*_{LD} across all values of L and D is 0.240 bits, or 30.5% of the *a priori* entropy. The corresponding numbers using a 12ms analysis window are 0.236 bits (30.0%), and using a 6ms window, 0.224 bits (28.4%). All infograms presented in this article are therefore based on a 20ms spectral window.

Infograms computed as described in equation 2 are displayed in figures 1 through 3. Infograms are titled in the format "landmark code:feature," where the landmark codes C, R, M, and P refer to closure, release, manner change, and pivot, respectively. Black denotes zero information, and white denotes information equal to I^*_{LD}.

4. DISCUSSION

The infograms in figures 1 through 3 show a distribution of information approximately as predicted by phonetic theory. Information about the features [sonorant] and [stiffvocalfolds] is located below 600Hz between closure and release, and information about [continuant] and [strident] is mostly above 2500Hz. Information about vowel features is mostly located in formant bands between -50ms and +50ms, with the notable exception of the feature [reduced], which seems to be cued primarily by timing information. The feature [continuant] at pivot landmarks also seems cued by timing information; this feature is used in this context to discriminate flaps from glides. The representation of place features such as [lips] is rather complicated, but there seems to be useful information present in all of the traditional cues: frication spectrum, formant onsets, and burst spectrum.

An infogram $I_{LD}(t,f)$ measures the mutual information between the distinctive feature and each spectral sample

individually, without consideration of the relationships between different spectral samples. There is a great deal of acoustic phonetic research which suggests that information about distinctive features is coded in the relationships between spectral amplitudes at different times and frequencies [6, 8], but if information is encoded only in spectral relationships, an infogram based only on $p(x, d)$ will not find it. In this sense, infograms are a good measurement of the amount of information about a distinctive feature which may be obtained *without* relational spectral cues.

The results in table 2 indicate that manner features such as [strident, continuant, sonorant] can often be identified based on the amplitude of a single well-chosen point in time-frequency space, while place features such as [lips, blade, anterior, distributed] are apparently difficult to identify without relational spectral cues. This result provides a novel explanation of consonant confusion studies, which have shown that the ability of listeners to identify place features degrades more rapidly in noise than their ability to identify manner features [4]: a single spectral sample with high SNR is often enough to recognize manner, but not place.

Future work will concentrate on measurements of the relative advantages of relational spectral cues over single spectral cues, as follows. The location of the most informative time-frequency coordinate for each distinctive feature will be fixed to the time and frequency discovered in section 3. With the most informative coordinate fixed, a "joint infogram" will be computed to estimate the joint mutual information which relates each spectral amplitude to the distinctive feature and to the previously selected "most informative coordinate." This procedure will be repeated to find optimal three-point and perhaps four-point acoustic correlates of each distinctive feature.

Distinctive Feature	H_{LD} (bits)	I^*_{LD} (bits, % of H_{LD})
strident	0.89	0.59 (67%)
fricative	0.95	0.53 (56%)
sonorant	0.81	0.45 (55%)
stop	0.91	0.43 (47%)
continuant	0.83	0.42 (51%)
high	0.96	0.41 (43%)
low	0.95	0.31 (33%)
CP	0.75	0.3 (41%)
consonantal	0.95	0.27 (29%)
back	0.95	0.25 (26%)
spreadglottis	0.78	0.24 (31%)
reduced	0.85	0.24 (28%)
stiffvocalfolds	0.97	0.21 (22%)
lateral	0.89	0.2 (22%)
rounded	0.83	0.19 (24%)
larynx	0.67	0.19 (29%)
grave	0.85	0.19 (22%)
ATR	0.64	0.17 (27%)
anterior	0.55	0.16 (30%)
blade	0.82	0.15 (18%)
constrictedglottis	0.96	0.13 (14%)
compact	0.77	0.12 (16%)
lips	0.7	0.11 (15%)
distributed	0.52	0.09 (17%)
syllabic	0.29	0.074 (25%)

Table 2. Information about each distinctive feature available from the best single spectral sample.

REFERENCES

[1] Andrew K. Halberstadt. *Heterogenous Acoustic Measurements and Multiple Classifiers for Speech Recognition.* PhD thesis, MIT, Cambridge, MA, Nov. 1998.

[2] Andrew K. Halberstadt and James R. Glass. Heterogeneous measurements and multiple classifiers for speech recognition. In *Proc. ICSLP*, Sydney, Australia, Nov. 1998.

[3] Sharlene A. Liu. Landmark detection for distinctive feature-based speech recognition. *J. Acoust. Soc. Am.*, 100(5):3417–3430, Nov. 1996.

[4] G. A. Miller and P. E. Nicely. Analysis of perceptual confusions among some english consonants. *J. Acoust. Soc. Am.*, 27:338–352, 1955.

[5] K. N. Stevens. Evidence for the role of acoustic boundaries in the perception of speech sounds. In Victoria A. Fromkin, editor, *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, pages 243–255. Academic Press, Orlando, Florida, 1985.

[6] K. N. Stevens. Relational properties as perceptual correlates of phonetic features. In *Proc. Eleventh Int. Conf. Phonetic Sciences*, volume 4, pages 352–356, Talinn, Estonia, 1987.

[7] K. N. Stevens, S. Y. Manuel, S. Shattuck-Hufnagel, and S. Liu. Implementation of a model for lexical access based on features. In *Proc. ICSLP*, volume 1, pages 499–502, Banff, Alberta, 1992.

[8] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J. Lang. Phoneme recognition using time-delay neural networks. *Trans. Acoust. Speech Sig. Proc.*, 37:328–339, 1989.

[9] Howard Yang, Sarel van Vuuren, and Hynek Hermansky. Relevancy of time-frequency features for phonetic classification measured by mutual information. In *Proc. ICASSP*, Phoenix, AZ, 1999.

[10] V.W. Zue, S. Seneff, and J. Glass. Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 9:351–356, 1990.

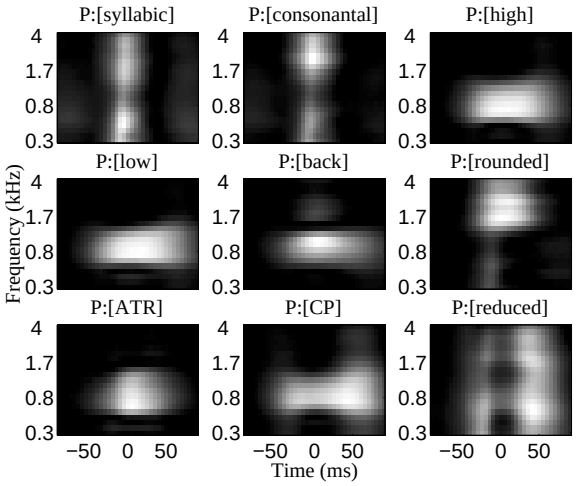

Figure 1. Features of syllabic nuclei ([+syllabic]), including vowels ([-consonantal]). Titles are in the form "landmark:feature," where landmark code is Pivot in this figure.

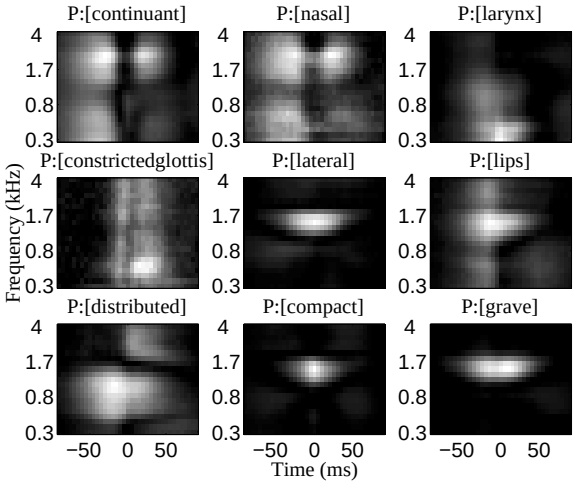

Figure 2. Place of articulation for flaps ([-continuant,+/-nasal]) and glides ([+continuant]). /h/ and /q/ are [+larynx] glides, distinguished by [constrictedglottis].

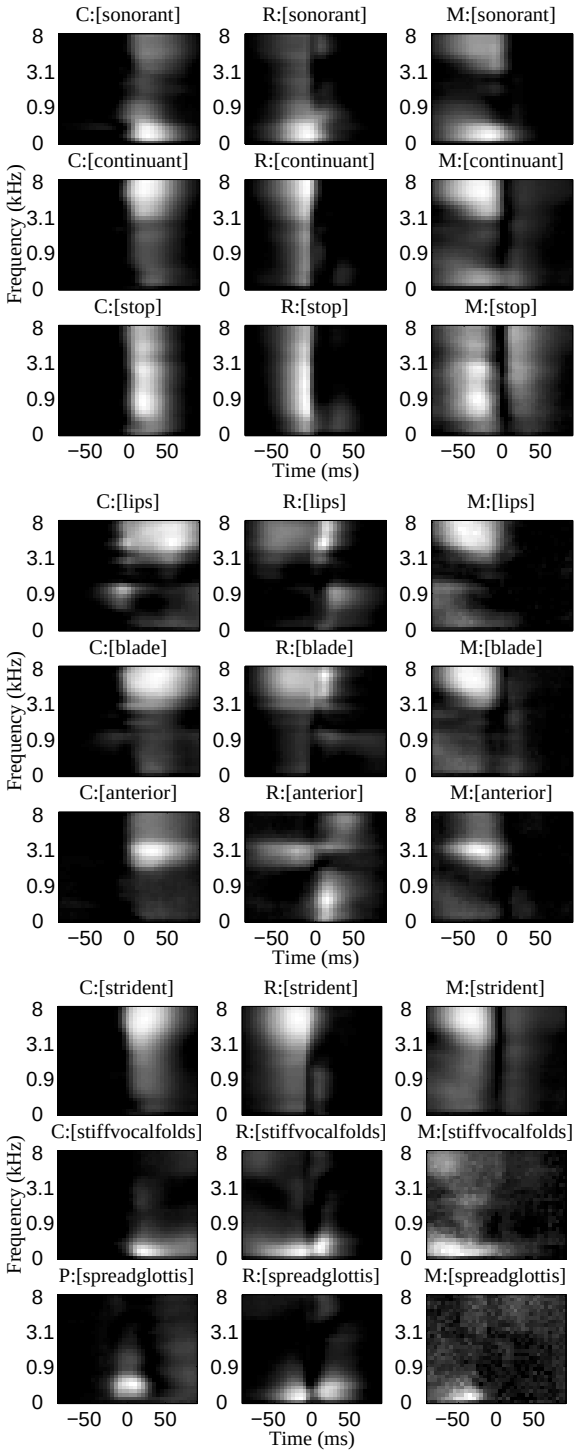

Figure 3. Features of consonants at closure, release, and manner-change. [stop] is redundant given [continuant] and [sonorant]. [+stiffvocalfolds] denotes obstruent devoicing. [+spreadglottis] marks aspirated stop releases, /h/ pivots, and devoiced schwa.