# ON PHASE PERCEPTION IN SPEECH

*Harald Pobloth and W. Bastiaan Kleijn*

Department of Speech, Music and Hearing
KTH (Royal Institute of Technology),
100 44 Stockholm, Sweden
Email: harald@speech.kth.se; bastiaan@speech.kth.se

**ABSTRACT**

In this paper we define perceptual phase capacity as the size of a codebook of phase spectra necessary to represent all possible phase spectra in a perceptually accurate manner. We determine the perceptual phase capacity for voiced speech. To this purpose, we use an auditory model which indicates if phase spectrum changes are audible or not. The correct performance of the model was adjusted and verified by listening tests. The perceptual phase capacity in low pitched speech is found to be much higher than it is for high pitched speech. Our results are consistent with the well known fact that speech coding schemes which preserve the phase accurately work better for male voices, while coders which put more weight on the amplitude spectrum of the speech signal result in better quality for female speech.

## 1. INTRODUCTION

It has been noted that phase information in periodic signals is far less important for high pitched signals than for low-pitch signals (e.g. [1]). In speech coding, it also has been observed that e.g. sinusoidal coders (which allocate relatively many bits to the description of the amplitude spectrum) perform better for female speakers while e.g. CELP (which represents phase information in a more precise way than sinusoidal coders) gives better results for male voices. (For voiced speech "phase" will refer to the phase spectrum of a single pitch cycle.) In this paper we quantify the perception of the phase, which is particularly useful for speech coding purposes.

In the 19th century, Ohm assumed that the human ear is phase deaf. During the last century, this assumption has been disproved by a number of workers; for an overview we refer to [1]. The fact that we can perceive phase information clearly suggests that the temporal distribution of the energy within a pitch cycle is of perceptual importance even when the amplitude spectrum of the signal is not effected. This temporal distribution of the signal energy is described by the phase spectrum of its Fourier transform.

Some work has been performed on phase perception in speech, but little has been quantitative. The results of Skoglund et al. [2] show that the audibility threshold for narrow-band noise decays significantly more between pitch pulses in low pitched vowels than it does for high pitched vowels. Their findings already give some evidence for the fact that a change of the time domain distribution of the signal energy has different perceptual significance for different pitch periodicities.

In this article, we describe a method which determines an upper bound for the bit allocation required to describe the phase spectrum of artificial voiced speech in a perceptually accurate manner. We refer to this upper bound as the *perceptual phase capacity*. It is defined as the size of a codebook of phase spectra necessary to represent all possible phase spectra in a perceptually accurate manner. This means there is at least one vector of phase values in this codebook which represents any arbitrary phase spectrum so that in the reconstruction no difference between the codebook generated signal and the original signal can be perceived (assuming the amplitude spectrum was not altered).

## 2. THE PERCEPTUAL PHASE CAPACITY

In this section, we develop a method to determine the perceptual phase capacity. Let us denote by $k_{\mathrm{opt}}$ the number of phase spectra in the smallest (i.e. optimal) codebook which can represent all phase spectra in a perceptually accurate manner. The perceptual phase capacity is then

$$C = \log_2(k_{\mathrm{opt}}) \; . \tag{1}$$

The entropy of the indices of the optimal codebook is defined as the perceptual phase entropy. If it is assumed that all phase spectra in the codebook appear with equal probability, then this entropy coincides with the perceptual phase capacity. In general, the probabilities of the codebook vectors will be unequal and the entropy will be less. Thus, the perceptual phase capacity forms an upper bound to the perceptual phase entropy.

For a strictly periodic signal, the phase spectrum can be obtained by computing the discrete Fourier transform of one pitch cycle. We denote the phase vector by $\vec{\phi} = (\phi_1, \phi_2, \cdots, \phi_n)^T$; it contains the phase values of the harmonic components of the complex DFT spectrum. This vector can be seen as a representation of a point in a hypercube of dimension $n$, where $n$ is the number of harmonics. Due to the periodicity of the phase the sides of the hypercube can be restricted to an edge length of $2\pi$.

To obtain our estimate, $k$, of the minimum number of codebook vectors $k_{\mathrm{opt}}$, we define the "perceptually equal region" around a point $\vec{\phi}$ in the phase-hypercube (PH) as the region which is associated with signals with no perceptual difference from that point, as illustrated in Figure 1.

The perceptually equal regions can be estimated using random sampling. Let $\vec{\phi}_i$ represent a random point and $V_i$ the associated perceptually equal region. Furthermore, let $V = (2\pi)^n$ denote the total volume of a unique region of the PH. We now sample $N_{\mathrm{test}}$ other random points in the PH and determine if these are within the perceptually equal region of $\vec{\phi}_i$ or not, using the auditory model described in section 3. We denote the number of points within the perceptually equal region of $\vec{\phi}_i$ by $\tilde{n}_i$. Then, ignoring statistical

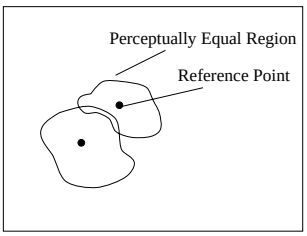


Figure 1: A possible arrangement of two points $\vec{\phi}$ in a two-dimensional PH and the perceptually equal regions around them.

fluctuations, we have that

$$\frac{V_i}{V} = \frac{\tilde{n}_i}{N_{\text{test}}}. \tag{2}$$

To obtain a reasonable estimate of $\frac{V_i}{V}$ we must make $N_{\text{test}}$ sufficiently large so as to obtain a number $\tilde{n}_i$ which is reliable.

The task is now to compute an estimate of the required codebook size from a set of relative volumes, $V_i/V$, for the perceptually equal regions. We assume that the chosen set of $V_i/V$ are representative for the entire PH. For our theory, we start with the assumption that the PH is already partitioned into an optimal set of perceptually equal regions, labeled $i$, each with a reconstruction point $\vec{\phi}_i$. We first select a random point in the PH to get a random region with index $I$, which is a random variable. The probability of selecting a particular region $i$ is just $\Pr(I = i) = V_i/V$. We then see that the total number of codebook entries, $k_{\text{opt}}$, can be written as

$$\begin{aligned} k_{\text{opt}} &= \sum_{i=1}^{i=k_{\text{opt}}} 1 = \sum_{i=1}^{i=k_{\text{opt}}} \frac{V_i}{V}\frac{V}{V_i} \\ &= \sum_{i=1}^{i=k_{\text{opt}}} \Pr(I=i)\frac{V}{V_i} \\ &= \mathrm{E}\left[\frac{V}{V_i}\right] \approx \langle \frac{N_{\text{test}}}{\tilde{n}_i} \rangle \,, \end{aligned} \tag{3}$$

where $\langle\cdot\rangle$ is the average of a set of measurements for random points $i$.

In a practical situation, we do not have a partitioning, as estimating the number of regions of the optimal partition was our goal. Instead, we use random points $\vec{\phi}_r$ as an approximation of the reconstruction values $\vec{\phi}_i$. That is, we assume that a random point $\vec{\phi}_r$ within a perceptually equal region $i$ has a perceptually equal region around it which is close to that around $\vec{\phi}_i$.

Alternatively, we can view the randomly selected point, as a reconstruction point in another, nearly optimal codebook, and we average the number of codebook entries additionally over these codebooks. It is reasonable to use equation 3 to estimate the minimum number of entries required in a codebook that can represent any phase spectrum in a perceptually accurate manner.

## 3. AUDITORY MODEL

To have an efficient method of deciding whether the difference between two signals with different phase spectra and equal amplitude spectra is audible or not, an auditory model is used. We use the functional path of the auditory image model (AIM) by Patterson et al. [3] to get a perception-related representation of the signals. To decide over the audibility of changes in this internal representation a simple decision criterion is used.

### 3.1. Auditory Image Model (AIM)

To get a representation of the activity in the auditory nerve fibers the neural activity pattern (NAP) [4] is used. The neural activity pattern is based on the basilar membrane motion using an adaptive thresholding mechanism [4]. The basilar membrane motion is the output of a gamma-tone filter bank [5] which simulates the spectral analysis of the inner ear. The filter bank used has $N_c$=24 channels ranging from 96 Hz to 3696 Hz which is the same range as in [1]. To obtain the NAPs and basilar membrane motions the AIM software [3] is used. The final step in the auditory model is a correlogram which is calculated similar to the correlogram in [6] except that the autocorrelation functions are normalized to the energy and not to the energy to the 3/4 power.

The periodicity of the signal allows the selection of a one pitch cycle segment as representative of the complete signal. The pitch cycle chosen to represent a signal is cut out of the middle of a 200 ms NAP as the NAP has reached a semi-stable state at this time instance. The fact that the decision criterion together with the correlogram allows the separation of the data from the listening tests in section 4.1, indicates that the correlogram is a good representation when the detection of phase changes is desired.

### 3.2. Decision Criterion

To decide about the detectability of differences in phase spectra, the correlogram is calculated for both the reference signal and a phase-altered test signal. From these two correlograms a decision criterion $\zeta$ is calculated:

$$\zeta = \frac{1}{\Delta\tau}\sum_{i=1}^{N_c}\int_{t=\tau-\Delta\tau/2}^{\tau+\Delta\tau/2} |\hat{c}_i(t) - c_i(t)|\,\mathrm{d}t \,, \tag{4}$$

where $\hat{c}_i$ is the autocorrelation function of the test signal's NAP in channel $i$ and $c_i$ is the corresponding autocorrelation function of the reference signal's NAP in channel $i$. $\tau$ is set in the interval $\tau \in [0, T)$ so that the expression $|\hat{c}_i(t) - c_i(t)|$ is maximized, where $T$ is the pitch interval. A $\Delta\tau = 1.25$ ms gave the optimal performance of the criterion. That means a sliding 1.25 ms window is positioned in each channel so that optimal detection is obtained. The normalization factor $1/\Delta\tau$ renders the criterion unitless. The pathway through the model is shown in figure 2.

When the signals used in the listening test (section 4.1) are used with the model, it can be seen that the criterion $\zeta$ is able to separate the signals for which no difference is perceived from the ones where there is an audible difference between reference and test signal. The threshold is set so that signal pairs which can be distinguished with a probability $\leq$50% are marked as perceptually equal and signals which can be distinguished with a probability $>$80% are marked perceptually different, see figure 3. The uncertainty between detection rates of 50% and 80% is not a drawback considering that listening tests with human subjects would give a high variation in this critical region as well. The final threshold value found in this manner is $\zeta_t = 1.6$. When the model is used to estimate the perceptual phase capacity, signal pairs resulting in $\zeta \leq \zeta_t$ are marked perceptually equal and pairs where $\zeta > \zeta_t$ are marked perceptually different.

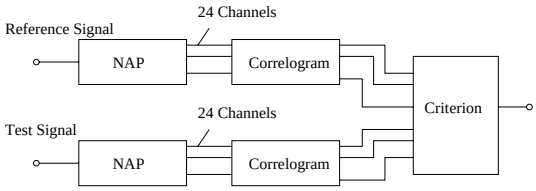


Figure 2: The auditory model applied to a pair of original and distorted speech signals.

## 4. EXPERIMENTS

### 4.1. Listening Tests

To ensure that our decision criterion is able to decide whether two signals with two random phase spectra are perceptually distinguishable, listening tests were performed. The listening tests were carried out in two steps. First an 1up-/ 2down-procedure described by Levitt [7] was used to approximate the point of 70.7% right responses to the 3 alternative forced choice (3AFC) described below. Then the method of constants [7] was used to get a better resolution of the psychometric function taken directly from the 3AFC responses without further statistical analysis.

Both procedures (the up-/ down-procedure and the method of constants) require a reference signal and a test signal which gradually drifts away from the reference signal. The reference signal had random phase $\vec{\phi}$ but the gradual introduction of distortion into the test signal can not be achieved by randomizing the phase spectrum. So the phase changes for the test signal were performed in a time shift invariant space where pure time shifts of the reference signal to the test signal are avoided. In this new domain a vector $\vec{w}$ with $||\vec{w}|| = 1$ defines the direction in which the phase changes. A scalar $\lambda_{tot}$ defines how far away the phase spectrum of the reference signal and the test signal are from each other. $\lambda_{tot}\vec{w}$ gives a complete description of the difference between the phase spectrum of the reference signal and the test signal.

In all tests, the following 3 alternative forced choice (3AFC) setup was used. For each decision, three sounds AXB were played to the subject binaurally. The level was set so that all sounds resulted in approximately 80 dB SPL in the ear channel of the subject. The first and last sound were the reference and the test signal in random order, the second sound was a random choice from these two sounds. The subject had the following three choices: X=A or X=B or X was not distinguishable from both of them. This is the same as the options: the first two sounds were equal, the second two sounds were equal or all sounds were equal. A user answer was marked "right" when the subject recognized the right pair (AX or XB) to be equal.

During the second stage of the listening test, a set of test signals with $\lambda_{tot}$ varied from 0 to $\lambda_{70.7} + \pi/2$ in steps of $\pi/4$ were generated for each reference signal. All signal pairs were played out to the subject $2 \cdot 20$ times in random order. $\lambda_{70.7}$ is the value for $\lambda_{tot}$ when the 1up- /2down-procedure indicated 70.7% right responses. In this configuration equal signals A and B were presented to the subject when $\lambda_{tot} = 0$. This was done to find out if the subject marks the signals to be equal in this listening situation which was found to be true. The output of the auditory model as a function of detectability can be seen in Figure 3. Detectability is defined here as the fraction of right answers to the 3 AFC procedure. Both tests were performed double blind with one subject,

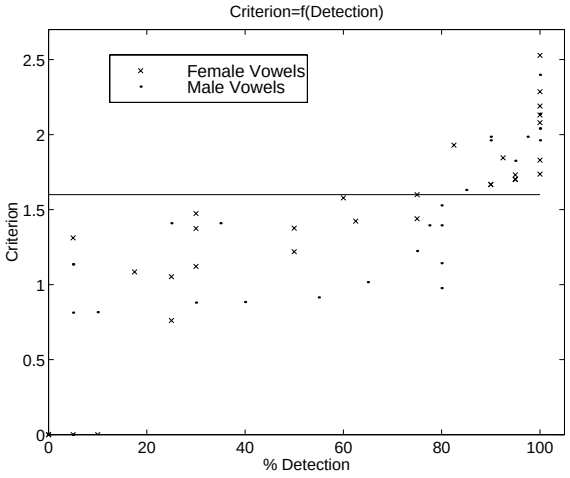


Figure 3: The criterion $\zeta$ as a function of detectability. The horizontal line marks the threshold

so that a good estimate of the subject's psychometric function was achieved.

### 4.2. Perceptual Phase Capacity Estimation

For the experiments, voiced segments from the utterance "She had your dark suit in greasy wash water all year." spoken by different speakers sampled at $f_s = 8$ kHz were selected from the TIMIT database. One pitch cycle of the signal is cut out of the "$a$" in "dark", "$i$" in "in" and "o" in water, see Table 1. Each of these pitch cycles was filtered through a linear prediction (LP) filter with the 10 LP coefficients optimized for this segment of the utterance and speaker. The residual signal was subject to the phase manipulations. The reason to use the residual domain was to reduce windowing effects which could occur when the pitch cycle is extracted from the original signal. To obtain signals which have sufficient relation to speech and a well defined phase spectrum the one pitch-cycle residuals were concatenated to 700 ms long signals. Finally, speech domain signals were obtained by feeding the residual signals through the LP synthesis filter after the phase manipulations.

By generating a number of random phase vectors $\vec{\phi}$ with independent and identically distributed (*iid*) samples of equal probability in the range of $\phi_i \in (-\pi, \pi]$, $i = 1, \cdots, n$, the entire phase-hypercube (PH) containing all possible phase spectra is covered with equal probability.

For the approximations presented here, we generated $N_{\mathrm{ref}} = 100$ reference points $\vec{\phi}_r$ for two vowels from two speakers, one male, one female. For these points $\vec{\phi}_r$ we found if there was a perceptual difference to $N_{\mathrm{test}}$ randomly chosen test points in the PH. $N_{\mathrm{test}}$ and $C$ for the different vowels are shown in Table 1.

The method of random sampling requires $\tilde{n}_i > 0$ for all $i = 1, \cdots, N_{\mathrm{ref}}$. Otherwise the approximation

$$C \approx \log_2\langle\frac{N_{\mathrm{test}}}{\tilde{n}_i}\rangle = \log_2\left(\frac{1}{N_{\mathrm{ref}}}\sum_{i=1}^{N_{\mathrm{ref}}}\frac{N_{\mathrm{test}}}{\tilde{n}_i}\right) \qquad (5)$$

does not give finite results. The number $N_{\mathrm{test}}$ used in the experiments was not high enough to obtain all $\tilde{n}_i > 0$. As the computa-

Table 1: Perceptual phase capacities for $N_{\mathrm{ref}} = 100$.

| Speaker, Vowel: | Pitch: Hz | $N_{\mathrm{test}}$: | $C$/Bit: |
|---|---|---|---|
| female, 1: $a$ | 216 | 10000 | 6.36 |
| female, 2: $i$ | 200 | 10000 | 6.38 |
| male, 1: $a$ | 114 | 92000 | 16.9 |
| male, 2: o | 93 | 68400 | 17.5 |

tional effort connected with producing a sufficient number of test points was very high, values obtained from thresholds $\zeta_t > 1.9$ were interpolated to the threshold value $\zeta_t = 1.6$ found in section 4.1. The values and the interpolation can be seen in figure 4. The figure shows also that our 3rd order polynomial interpolation coincides well with the data points at threshold $\zeta_t = 1.6$ when it is applied to the female vowel capacity $C$.

## 5. RESULTS

The values in Table 1 confirm the assumption that the perceptual phase capacity is lower for high pitched (female) speech than it is for low pitched (male) speech. When one codebook vector index has to be transmitted every pitch cycle, which is reasonable when it is assumed that the vectors are exactly one pitch cycle long, the transmission rate needed is

$$r = \frac{C}{T} = \left[\frac{\mathrm{Bit}}{\mathrm{s}}\right] . \tag{6}$$

To keep this rate constant $C$ has to be proportional to the pitch period $T$. Considering the values $C \approx 6.4$ bits for $T \approx 5$ ms and $C \approx 17$ bits for $T \approx 10$ ms even the rate $r$ increases with increasing pitch period. The values for $C$ show a clear dependency of the perceptual phase capacity on the pitch period. However, from what is known about simultaneous masking it is clear that $C$ must depend on the amplitude spectrum as well.

Our method is a simple way of estimating the perceptual phase capacity $C$. The assumptions about the geometry of the perceptually equal region are kept to a minimum. It is, for example, not necessary that the perceptually equal region is a continuous hyperspace.

For low pitched vowels $C$ is so high that the computational effort required to do the sampling with the necessary number of test points $N_{\mathrm{test}}$, becomes almost unrealistic with the current implementation of the sampling procedure. Third order polynomial interpolation was able to overcome this problem as all values of $C$ obtained from the sampling procedure are approximated in a sufficient manner. The steep slope of $C = f(\zeta_t)$ in figure 4 might indicate that the auditory model is too sensitive in some regions.

## 6. CONCLUSIONS

Taking into account the relatively low attention the phase spectrum historically has received in speech coding and psychoacoustic research, the perceptual phase capacity was much higher than expected. That means humans are able to distinguish between different phase spectra much better than often is assumed.

The current method provides values of the perceptual phase capacity which can help when decisions about the precision of phase reconstruction in speech coders have to be made. It will

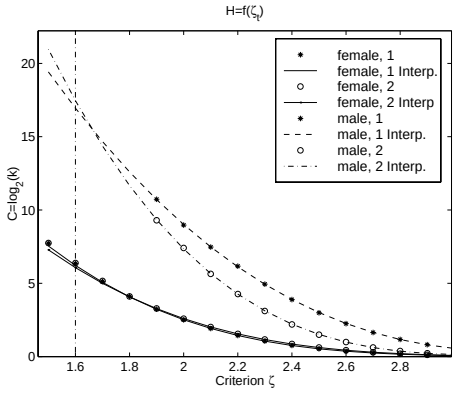


Figure 4: The perceptual phase capacity as a function of the threshold value $\zeta_t$. The curves marked as Interp. are the third order polynomial interpolations from the existing data points. The vertical line shows the experimental observed threshold $\zeta_t = 1.6$.

be useful to extend the present results to other speech sounds. The perceptual phase capacity is particularly interesting for variable rate coders which can adjust the method of phase encoding as a function of pitch period.

Statistics of typical speech phase spectra will also provide insight into how different the perceptual phase capacity is from the perceptual phase entropy. As voiced speech has certain pattern statistics it is possible that the entropy is far lower than the capacity.

Our results show that the pitch-dependent performance of coding schemes like CELP or sinusoidal coders can partly be explained by the pitch dependency of the perceptual phase capacity. Considering the results in Table 1, it is most likely that current speech coders introduce phase distortions well above the audible threshold.

## 7. REFERENCES


[1] R. D. Patterson, "A pulse ribbon model of monaural phase perception", *J. Acoust. Soc. Am.*, vol. 82, pp. 1560–1586, 1987.

[2] J. Skoglund, W. B. Kleijn, and P. Hedelin, "Audibility of pitch-synchronously modulated noise", *in Speech Coding Workshop*, Pocono Manor, PA, 1997.

[3] R. D. Patterson, M. H. Allerhand, and C. Giguère, "Time-domain modeling of periphal auditory processing: A modular architecture software platform", *J. Acoust. Soc. Am.*, vol. 98, pp. 1890 – 1894, October 1995.

[4] J. Holdsworth and R. D. Patterson, "Analysis of waveforms", Technical Report UK Patent No. GB 2-234-078-B, UK Patent Office, London, 1993.

[5] R. D. Patterson, "The sound of a sinusoid: Spectral models", *J. Acoust. Soc. Am.*, vol. 96, pp. 1409 – 1418, 1994.

[6] M. Slaney and R. F. Lyon, "A perceptual pitch detector", *in Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, vol. 1, pp. 357 – 360, Albuquerque, 1990.

[7] H. Levitt, "Transformed up-down methods in psychoacoustics", *J. Acoust. Soc. Am.*, vol. 49, pp. 467 – 477, 1971.