

Richard J. Fabbri
Scarce Ideas Co.
Stamford, CT, USA

**Presented at
the 94th Convention
1993 March 16-19
Berlin**



AES

This preprint has been reproduced from the author's advance manuscript, without editing, corrections or consideration by the Review Board. The AES takes no responsibility for the contents.

Additional preprints may be obtained by sending request and remittance to the Audio Engineering Society, 60 East 42nd Street, New York, New York 10165, USA.

All rights reserved. Reproduction of this preprint, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

AN AUDIO ENGINEERING SOCIETY PREPRINT

Phoneme Recognition During a Cocktail Party

Richard J. Fabbri
Scarce Ideas
rfabbri@earthlink.net

ABSTRACT

Today's automatic speech recognition has reached fundamental limits, i.e., Isolated (single) Speaker and Limited Vocabulary. Wave Precedence, Interaural Delay and Masking are combined in the time domain using simple logic, fractal geometry and neural networks to reach Star Trek goals, i.e., Cocktail Party (multiple-speaker) Processing and Phoneme Recognition (Unlimited Vocabulary). Demonstrations illustrate these synergistic principles. The following is a recent edit (Sept 1999) of an AES preprint (and talk) delivered at Berlin in March 1993.

INTRODUCTION

Language text is found in documents such as this AES preprint [1]. While scanning this text your brain converts printed words to the sounds of spoken language: an inner voice reads to you. Text, a media construct, is converted to speech by reading. Speech is the essence of language. We think in, and communicate using, speech. How would artificial speech communication work, i.e., just as text symbols are computed by this word processor, are there spoken symbols a speech processor might recognize and compute? What is the shape of our spoken symbols: what shape (encoding) does the brain use? Recognition of spoken symbols would allow the typical Star Trek conversation:

Captain ..."Computer, where's the nearest star base?"...

Computer ..."*Sector H345, in star system Riga, sir.*"...

Your mind heard the captain and computer speak while you read the text. But, on a ship's noisy bridge, the computer must first *separate* then *detect* sound coming from the captain's direction and then recognize that sound as speech: How do we do this?

Phonology, the study of language sound patterns and prosody, centers on the concept of phonemes. As a common example, all dictionaries introduce and define an International Phonetic Alphabet (IPA). The IPA is a set of written symbols representing a set of "international" sounds called phonemes. IPA symbols are acoustically defined by pronounced segments of commonly used words. Dictionaries then indicate accepted word pronunciation by acoustic sequences of IPA symbols.

The IPA, and similar systems describing the sounds of a spoken language, have a common problem: language is spoken by individuals. Men, women, boys, girls, foreigners speak unique *sounds* in order to *form* the words they speak. Even though fatigue, distress, etc. may further change pronunciation, we display a robust ability to understand all speakers, i.e., our awareness of pronunciation must be distinct from our recognition of speech.

Furthermore, hearing accommodates the frequencies of all speakers. Anyone who has answered a question, posed by a complete stranger, understands this accommodation. Or, noted that children (having higher pitch) learn to speak by listening to adults normally having lower pitch!

Thus, our sense of hearing seems to *ignore* frequency. However, all speech researchers presently use frequency *analysis* to measure and classify unique phoneme speech patterns. Actually, "frequency" is a carrier-of-meaning at levels higher than phoneme recognition (prosody, etc.). But, frequency *analysis* has distracted us from the larger task of recognizing *speech information*.

Fourier analysis (averaging and linear superposition) ignores the bursty, time-encoded structure of speech (see Fig.1). However, all speech *structure* implies information. Just as Fourier analysis finds model speech spectra, time domain patterns, obvious to the eye prior to Fourier transformation, reveal a time sequential structure to spoken sound. When appropriately displayed, these acoustic sequences reveal their geometric form as fractal-object events synchronized to our real-time speech. Specifically, binaural time-cues (10 μ sec region) [2] sort acoustic sources within the horizontal plane (see Fig.2) and, are further used to recognize the separated sources by their geometric form. These essential 10 μ sec cues are silenced by the averaging process inherent to spectral analysis. However, (described later) bursty speech patterns can be isolated and defined (... "spoken symbols") using a new, time-based, binaural process and a SOM (neural net) Self-Organized by the statistics of these patterns.

The Cocktail Party Effect (CPE) is solved by this binaural scheme and allows ensuing analysis to focus on acoustic *sources*. Similar to binocular vision, binaural hearing gives acoustic perspective, i.e., speakers and environmental sounds are localized (as are their reflections!). Although early reflections (<35msec) are silenced by Wave Precedence, reverberant fractals (p.4, col. 1) are unconsciously used to estimate source distance. Contrasting, when listening with one ear (as during a telephone call) sources are hopelessly mixed and all reflections are heard. More important to speech recognition, think of that telephone call as representing a single-point mic that permanently mixes voices (and reflections) with the environment ... this represents the normal (*expected*) input for state-of-the-art ASR systems!

THE COCKTAIL PARTY EFFECT

The first challenge of artificial speech recognition is isolating the speech signal intended for analysis. However, if the larger goal is to solve the CPE, *all signals* must be isolated at once.

The first challenge is usually attempted by close-miking the intended speaker; this assumes the speaker's signal will overcome the level of *nearby* speakers and environmental sounds usually called "noise." This is an algebraic summation characterized by a hopefully large signal and (equally hopeful) small noise; the audio engineer assumes close miking produces both. However, one source of problem occurs when the speaker is silent and the "noise" sources are not. This condition illustrates the fact that *other* sources are *equally* weighted and not *excluded* from analysis. The usual engineering fix amounts to a manual (or automatic)

microphone "standby" during speaker silence and/or a directional microphone.

However, if our goal is to solve the CPE, other speakers are not noise: they are the source of other speech signals! The single speaker case is a subclass of the CPE and is "solved" today by close miking ("localizing") the single speaker. For the CPE, other speakers and environmental sounds are not considered noise; each is an acoustic source, i.e., "noise" transforms to a group of localized, acoustic sources.

How do we simultaneously localize all speakers at a cocktail party? The first clue is we use more than one microphone. We normally have and use two ears. What class of information do our ears deliver, and our brains seek, that together equals more than either ear individually? This is a prime example of synergy, i.e., each ear has the same algebraic summation "problem" of a single microphone yet, together, the hearing process *unmixes* this pair of summed signals. There will be an audio demonstration of our binaural unmixing, localizing abilities at the end of this paper. Note the clarifications just employed: unmixing has been linked to localizing while noise has been transformed to isolated signals at those locations.

The key question becomes: which acousto-mechanical properties of the head provide our directional encoding mechanism? Encoding is the key word. Our physical head must be the source of the cues our directional decoding uses to localize speakers. If we resort to frequency analysis, head structures act as filters. However, the decomposition of incoming acoustic energy into (acoustic) sine waves does not indicate how to choose subsets of those sine waves to reconstruct the

original waveforms of the acoustic speech sources. Sources must be separated prior-to any kind of analysis otherwise analysis must separate **and** analyze sources. Even "directional" frequency response (as a logic to locate sources) needs an additional mechanism to decide how to identify the *particular* frequencies of a *particular* source to render (perceived) source timbre.

What is it about hearing that can easily be identified with localization? Is there a psychoacoustic effect that easily demonstrates how we localize? There is; the effect is called Wave Precedence (WP). The "acoustics" of concert halls could never have been invented without the concept of WP. To state it simply, WP means the initial acoustic wave (reaching our ears) establishes the *perceived* direction to an acoustic source. "Early" reflections (reflections arriving approx 600µsec to 35msec later) do not *change* that perception. More importantly, the arrival direction of early reflections is consciously ignored. This point is discussed more fully later: early reflections are **unconsciously** used to estimate the distance to acoustic sources using the Law of Cosines and, add to perceived source volume.

The principle that explains WP also defines what we mean by acoustic localization. The maximum interaural time delay is approx 600µsec: a signal arriving directly from our side takes approx 600µsec to propagate around the head and be heard at the opposite ear. A signal arriving from the Median plane has *no* interaural delay; the signal reaches each ear at the same time (see Fig. 3 for a graphic of the Median plane). The range of interaural (binaural) delay is thus 0 to 600µsec. This varies with size of head, but the

range is always 0 to a maximum 100's of μsec .

The brain uses this crucial fact to localize acoustic sources, i.e., as soon as the brain perceives an event in one ear it looks to find that **same** event in the *other* ear. A maximum interaural delay guarantees this will always occur in natural situations. During WP the initial wave front arrives and we immediately compute the (source) arrival direction from the measured interaural delay (Fig. 2).

The first reflections arrive and, are detected as such because they exceed the max binaural delay of $600\mu\text{sec}$. Generally, sounds are tagged as reflections if binaurally similar to an earlier wave front, i.e., are rejected as coming from a *source* direction. The brain does not report these early reflections as "sources" to our conscious mind, i.e., we do not perceive early reflections as "echoes": reflections are consciously ignored as sources until they exceed an echo threshold ... at that point we (consciously) perceive (source) echoes. Early reflections ($<35\text{msec}$) are now termed "reverberant fractals".

Reverberant fractals give us a perceived sense of room size and shape. We are not conscious of the trigonometric (Law of Cosines) projections in progress, i.e., a neural net map uses the delay and arrival direction of reverberant fractals to sense the distance of an acoustic source and, at a higher level, synthesizes a spatial map of the room (it's reflecting surfaces). We sense the hard-soft, cold-warm qualities of "room reverberation" as a consequence of reverberant fractal calculations and reflective (echo) perception.

If you block one ear you'll hear the early reflections silenced during binaural hearing. Early in mammalian evolution, the ability to perceive the direction to sound sources (by silencing early reflections) had survival value: we knew which way to run to either kill our food or avoid being eaten. Imagine not knowing the direction toward the subtle sounds (or the roar) of a predator! Today we, the descendants of our early survivors, have applied WP to the problems of concert hall design. Loud speakers near our concert seats are purposely delayed from $600\mu\text{sec}$ to 35msec beyond the arrival of **direct** stage sound ... we perceive sound as coming from the direction of the stage yet, we *hear* the *volume* of the *local* loudspeakers!

To summarize, our key fact has been the existence of a maximum interaural delay (approx $600\mu\text{sec}$). Using real-time, interaural source delays we unconsciously draw a map of acoustic source(s) direction(s) (interaural delays $<600\mu\text{sec}$). Via reflecting surfaces (reverberant fractals $>600\mu\text{sec}$ and $<35\text{msec}$) we apply a scale to those source directions and produce a spatial map of our environment. However, this reverberant process is frustrated during a common activity such as a walk on the beach. That is, even though we can perceive the direction to others on that beach (interaural delays $<600\mu\text{sec}$), we find it hard to tell their distance, i.e., without reflecting surfaces WP finds no reverberant fractals and therefore can not scale (estimate) source distances.

Through many psychoacoustic experiments it is well known that our minimum "audible" angle (MAA) [2] is in the range of one (1) to two (2) spatial degrees. It is a simple trigonometric exercise to derive a relationship

between a measured interaural delay and the spatial angle (to an acoustic source) producing that interaural delay. Figure 2 is a plot of interaural delay versus source angle, graphed by an Excel spreadsheet, using this simple trigonometry. As shown in Fig. 2, a spatial resolution of one (1) degree (MAA) equates to a time delay resolution of $10\mu\text{sec}$.

3-DIMENSIONAL HEARING

How do we build a 3-D spatial map? We don't live in Flatland, i.e., besides angles in the horizontal plane (raw interaural delay) we also perceive elevation angle in both the front/back and left/right directions. To explain these 3-D acutities it is necessary to return to the earlier question of the acousto-mechanical properties of the head and understand the directional encoding mechanism these properties afford.

If we test our hearing with sine waves, we learn how we hear sine waves. But, we speak and hear bursty, acoustic cues well within the window times of Fast Fourier Transforms (FFT); as mentioned above, we sense an MAA equivalent to $10\mu\text{sec}$. Consequently, our binaural hearing must resolve $10\mu\text{sec}$ intervals. However, FFT windows of 10msec (see Fig.1 tic-marks) are near the lower limit of FFT time resolution because windows smaller than 10msec imply a spectral resolution grainier than 100Hz . In fact, a $10\mu\text{sec}$ FFT analysis window implies a spectral resolution of 100KHz , well *outside* the audio range! The acousto-mechanical properties of the head, which encode these $10\mu\text{sec}$ directional cues, must therefore operate in the time domain.

There are obvious time domain mechanisms when one considers the propagation of acoustic waves along

the surface of the head and through the various flesh, muscle and bone structures. Ultrasonic fetal monitoring stands as a ready example of the latter and uses time correlation techniques to investigate sub-surface structures.

The surface of the human head exploits the time delay aspect of surface acoustic wave propagation to generate time cues that depend on direction of wave arrival. The 3-D structure of the head is an encoding engine: it delivers time cues used during binaural hearing (**decoding**). Thus, 3-D acoustic perception is a consequence of our ability to process interaural delays at $10\mu\text{sec}$ resolution and decodes these encoded time cues to recreate 3-D wave front arrival direction.

The 3-D structure of the head is shown in Fig. 3 which also indicates the three principle reference planes called Horizontal (HP), Frontal (FP) and Median (MP). We are usually referring to the HP when discussing our ability to perceive different speakers at a cocktail party. The FP and MP are used to discuss elevations above and below the HP with reference to left/right (FP) and front/back (MP).

The intersection of the head with the HP is symmetric about the MP. Signals arriving within the HP are thus assigned the left or right side of the MP via the algebraic sign of their interaural time delay. Since the head's intersection with the HP is asymmetric about the FP, signals arriving within the HP from the front or back will be time cue encoded uniquely dependent on front/back arrival direction. The pinnae provide additional time cue encoding for the shorter wavelengths in the form of unique left/right and front/back delays cued to arrival direction. A similar discussion of time cue encoding is

possible for signals arriving in the FP since the intersection of the head with the FP is also symmetric about the MP. Since the intersection with FP is asymmetric about the HP, signals arriving within the FP from the up or down direction will be time cue encoded uniquely dependent on up/down arrival direction. The pinnae provide additional time cue encoding for the shorter wavelengths in the form of unique left/right and up/down delays cued to arrival direction.

The Median plane is the most important example of these surface acoustic delay encodings and requires a more rigorous discussion that takes us into the fractal-object events that represent spoken symbols.

First, it's clear (see Fig.3) the head's intersection with MP is the so-called "profile" of the head. The profile has no symmetries, i.e., it is asymmetric about both the FP and HP and thus must provide time-cue encoding for signals arriving in the MP from the front/back **and** up/down directions. Consequently, signals arriving from within the MP are not encoded with any left vs. right cues. That is, signals arriving from the MP exhibit no interaural delay differences. In the HP and FP it's clear that interaural delay polarity helps us separate left from right and, via a max interaural delay (<600 μ sec), allows WP to decide source direction by detecting a first arrival and hunting it's successive reflections.

Once again, in the Median Plane, there is no interaural delay. But, zero interaural delay still satisfies WP (<600 μ sec), i.e., the signal is simply coming from a source in the MP, and reverberant fractals of that source will most likely arrive from directions outside the MP and determine the distance to

the MP source. However, on that initial (MP) arriving wave, what do MP delay cues provide that is used during binaural hearing since interaural delay is 0 μ sec in the MP? Furthermore, since an MP source under anechoic conditions never generates reflections, (head) profile asymmetries must generate the only front/back and up/down delays. To use this data, one must first understand how this new time domain processing works then, apply it to the specific case of 0 μ sec interaural delay in the MP.

Think of a spreadsheet whose rows and columns are labeled by time parameters. Columns are labeled from -600 μ sec to +600 μ sec with a min column resolution of 10 μ sec, i.e., there are up to 121 columns. A central column is labeled "0 μ sec" and represents the whole of the Median plane. Particular columns contain source waveform information at specific interaural delays such as -360 μ sec, +520 μ sec, etc. These column headings represent angular directions in the HP as decoded by the relationship plotted in Fig. 2. The information found while going down the rows of a given column represents waveform information coming from a direction in the Horizontal Plane implied by the column label. The range of column labels span +/-600 μ sec about the Median plane.

This +/-600 μ sec range was chosen to agree with the Wave Precedence effect, i.e., interaural delays longer than 600 μ sec cannot arrive from a single ("source") direction. Stated differently, interaural delays up to the maximum (600 μ sec) can indicate a common source in a *particular* HP direction. The interaural columns of the spreadsheet are a first step in solving the CPE. This binaural spreadsheet is an example of what Daniel Dennett [3] calls Multiple

Drafts, i.e., "a competition among multiple patterns of mental activity propagating within the brain". Column data is a time sequential record of acoustic activity in either (or both) of the two vertical half-planes that pass through the (two!) HP angles that *ambiguously* represent each column (see Fig. 4). This half-plane "ambiguity" will soon be discussed and represents a well known psychoacoustic effect.

Sequential row information is plotted in a multidimensional space where geometric structure indicates novel aspects of incoming acoustic signals. These objects are the spoken symbols referenced in the introduction, i.e., they have unique geometric properties linked to the sounds of a given language. And, just as we recognize speech independent of pronunciation, certain geometric properties of these objects are independent of pronunciation. Additionally, geometric "distortions" encode angular information concerning "front/back" and "up/down."

Decoding these "distortions" solves the ambiguity previously mentioned, i.e., allows the choice of a single vertical half-plane (front/back decision) and a unique elevation angle (up/down decision) within that (decoded) half-plane.

A SOM type neural net, organized by the statistics of these acoustic objects, combs real-time data for geometric properties representing the primary sounds ("phonemes") of a given language. The SOM identifies object type (spoken symbol) and decodes the front/back - up/down cues.

All spreadsheet columns are independent: each column generates (perceived) fractal objects over time. Thus, there are up to 121 independent,

multi-dimensional spaces each linked to a spreadsheet column and each potentially evolving objects at the average rate of 12 per second, i.e., 2.5 average words per second times 5 objects (phonemes) per average word. These acoustic objects are plotted multi-dimensionally (using simple operations) while the SOM is time-shared by sequencing through active acoustic spaces (columns). This data is processed in real-time, i.e., in lock step with the evolution of sources in all columns. This is a full, dynamic solution to the Cocktail Party Effect and is self-clocked by data (cues).

Let's return to the Median Plane, i.e., zero interaural delay. "Zero" means the MP is formed by the two vertical half-planes constituting the front/back decision discussed above! As with any spreadsheet column, we begin to resolve MP direction by decoding (choosing) a vertical half-plane (front/back decision). Within that chosen half-plane, the decode of elevation angle (up/down decision) completes the source direction decode. Thus, the central spreadsheet column (MP) is processed with the same SOM neural net used by all *other* columns!

However, as with all psychoacoustic effects, resolution in the MP is limited during simultaneous signals (masking) whether originating in the MP or not. Additionally, in this new analysis, masking is itself a time domain effect.

1 Originally published (given) at the 94th AES Convention, Berlin, Mar 16-19, 1993.

2 Harris, D J: "Psychoacoustics", The Bobbs-Merrill Company, Inc., pp 73-79 (1974).

3 Dennett, Daniel C: "Consciousness Explained", Little, Brown and Company (1991).

The Bursty Nature of Speech

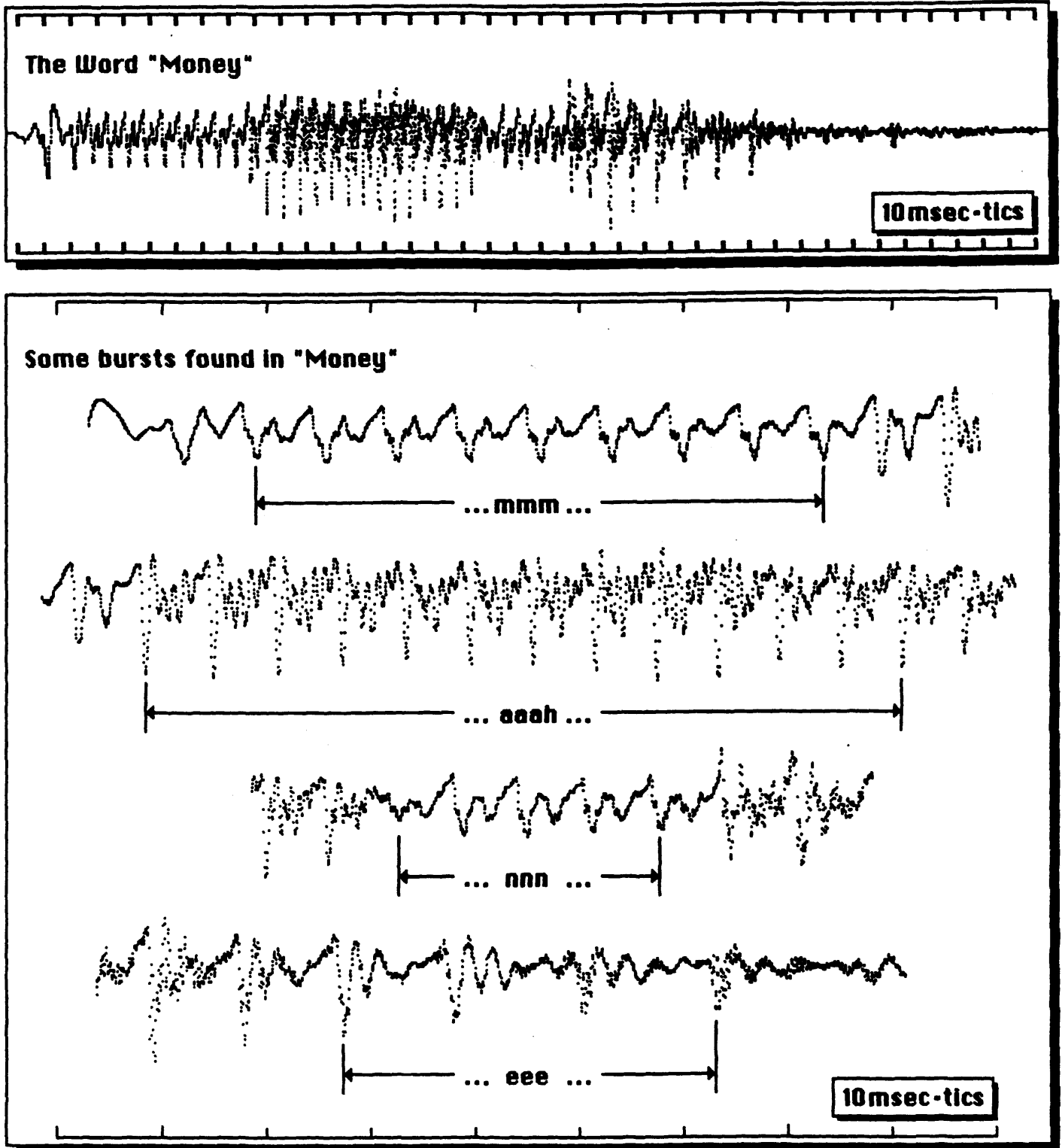


Figure 1

Direct Interaural Delay (20cm spacing) vs Azimuth Angle to Source

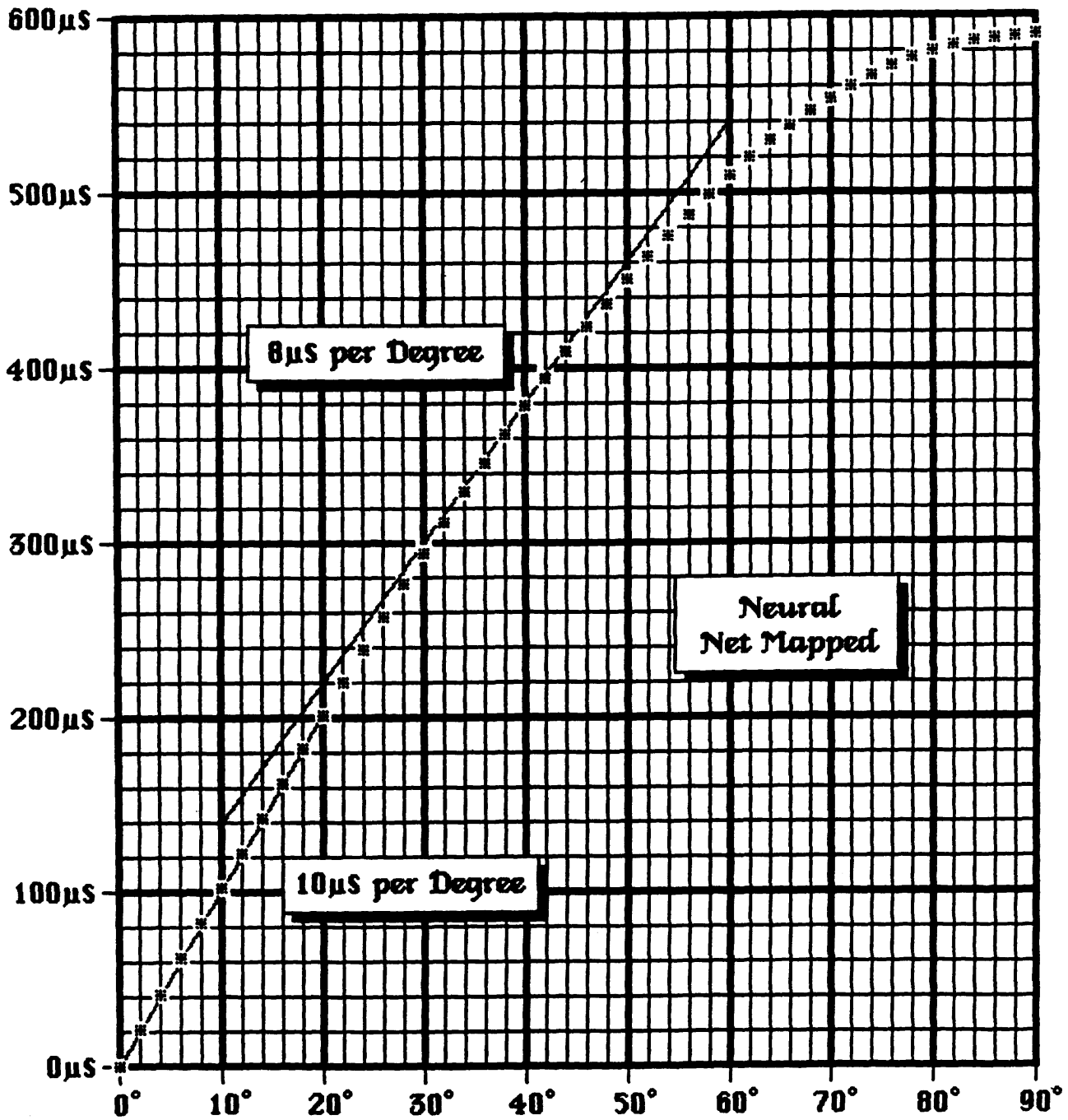


Figure 2

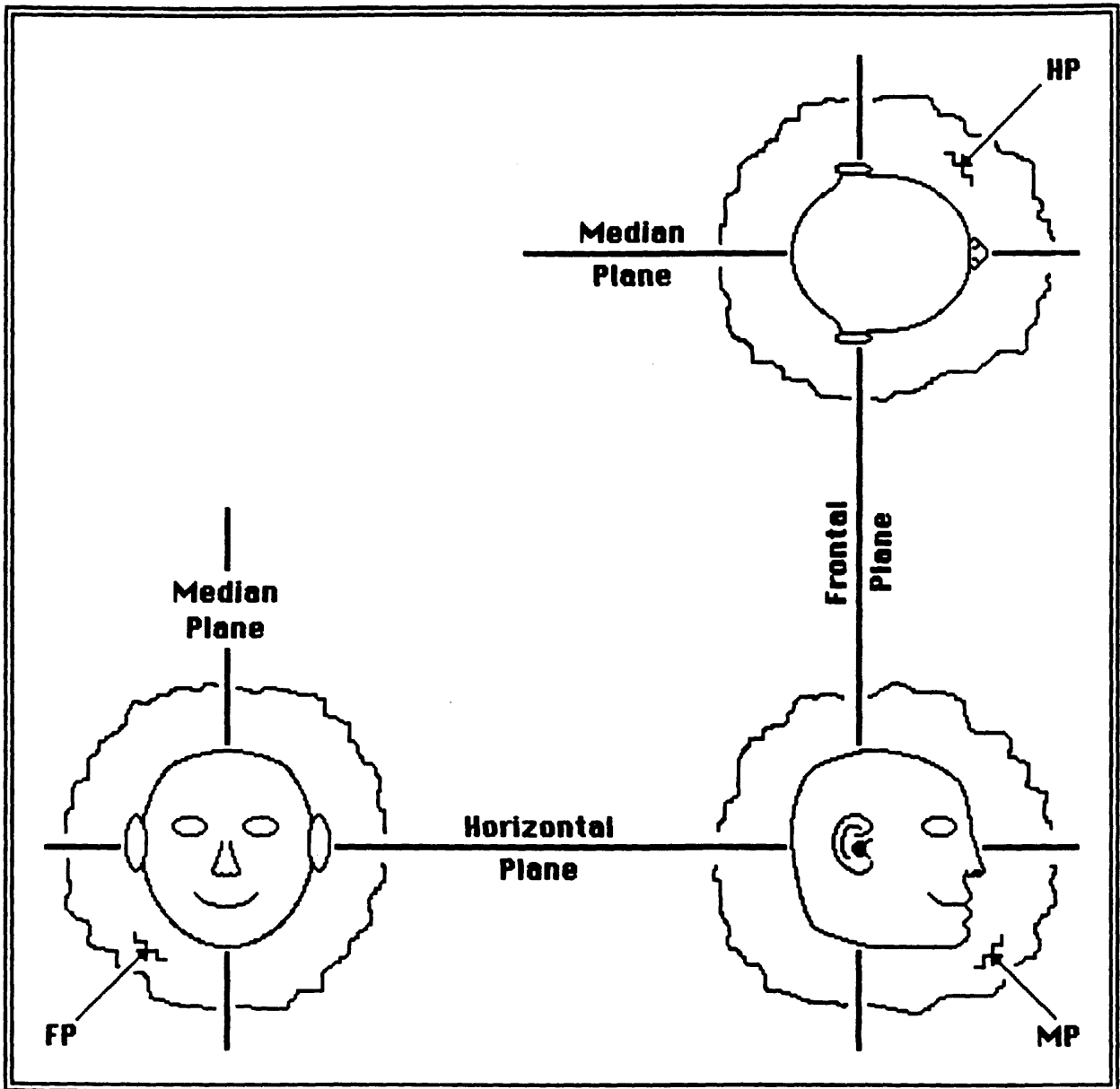


Figure 3: Head Reference Planes
The head is symmetrical within the Horizontal and Frontal planes, but asymmetrical in the Median plane.

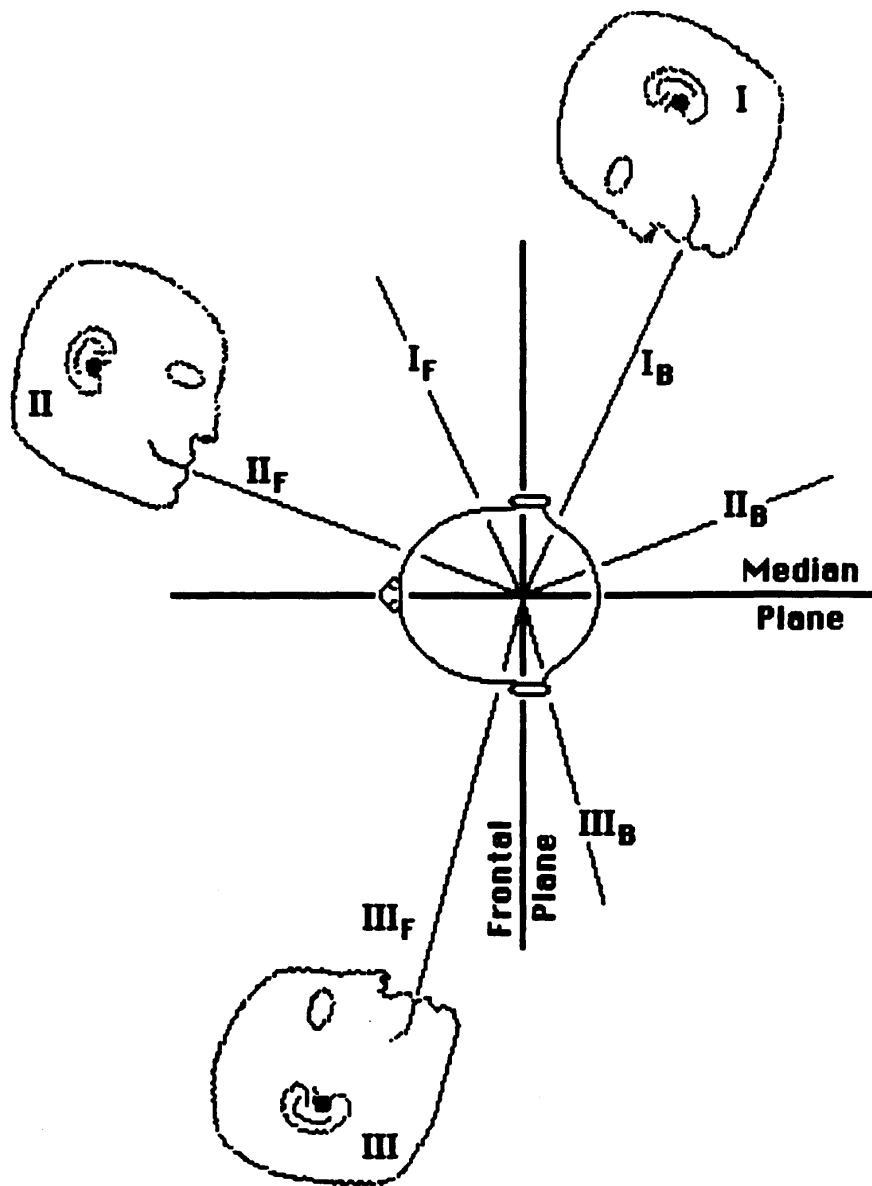


Figure 4: Vertical Half-Planes (F/B)

Two vertical half-planes correspond to a single spreadsheet column label (two HP angles)