INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Mapping Phonemes to Acoustic Symbols and Codes Using Synchrony in Speech Modulation Vectors Estimated by the Travellingwave Filter Bank

Ashwin Rao

Travellingwave, Seattle, WA, USA
ashwin@travellingwave.com

Abstract

A hybrid vector representation for speech resonances is defined using the modulation model and the sum of sinusoids model. An adaptive filter bank, whose channels utilize resonance localized modulation tracking, to robustly estimate temporal variations in these vectors, is then presented. The synchrony in modulations, within and across resonance channels, is subsequently used to derive acoustic symbols and codes that map fundamental units of languages, phonemes. Such an acoustic-phonetic mapping has never been demonstrated before. It has potential applications in speech recognition and voice analytics.

Index Terms: acoustic symbols, acoustic codes, acoustic cues, the speech code, phoneme mapping, symbolic representation

1. Introduction

Big-data systems that are currently used in applications like speech recognition [1] lack human-like performance and efficiency - their accuracy is susceptible to model mismatch [2, 3], they fail to provide reliable feedback for error-correction [4, 1], and they are very expensive to develop and deploy [1].

To address these problems, research on finding new acoustic cues in speech, which better map phonemes, has been underway for over a century [5, 6, 7, 8]. Many of these approaches are motivated by the way humans recognize phonemes, followed by syllables, words, sentences, and meaning [9].

Major strides have been made by Fant [6], Liberman [10], Stevens [11], Allen [12], and others [13, 14]. Their speech analysis experiments primarily rely on acoustic features estimated using the spectrogram [15], the linear prediction spectrum [16], and auditory filter banks [17, 12].

Unfortunately, successful mapping of phonemes has not been possible yet, due to a) high variability of existing speech features across speakers, phoneme context, and noise [6, 14], and b) limitations of time-frequency analysis tools [15] to jointly model phoneme transitions and resonances [14].

This paper introduces *three new concepts* for acoustic-phonetic mapping. The first, called modulation vector, is a hybrid representation for speech resonances that combines features from sinusoidal models [18, 19] and a generalized modulation model [20, 21, 22]. The second is an adaptive filter bank that improves upon the Rao-Kumaresan algorithm [22]; which was modified by Mustafa and Bruce in [23]. Specifically, it addresses problems in [22, 23] associated with complex-valued signals, frequency tracking errors, and filter instability. Additionally, it employs resonance localization to track modulation vectors in speech; instead of tracking formants as in [24, 25, 26, 27, 23], or modulated components (envelope and positive instantaneous frequency) as in [21, 22], or individual frequency components as in [28, 29]. Finally, the third concept utilizes synchrony in modulation vectors, within and across subbands, for mapping phonemes to acoustic symbols and codes.

In the remaining sections, modulation vector is defined in section 2, the adaptive filter bank is described in section 3, phoneme mapping using synchrony is derived in section 4, and simulation results are presented in section 5; discussions and conclusion follow in sections 6 and 7 respectively.

2. Modulation Vector

In [21, 22], the k-th resonance in a speech signal, $s[n]$, was expressed using the product of elementary signals [20] as

$$s_k[n] = a_{ck} e^{j2\pi f_{ck} n} e^{\alpha_k[n] + j\hat{\alpha}_k[n]} e^{\beta_k[n] - j\hat{\beta}_k[n]} , \qquad (1)$$

where n is the time sample, a_{ck} is the carrier amplitude, and f_{ck} is the carrier frequency. $\alpha_k[n]$ and $\beta_k[n]$ are details in modulations around f_{ck}; hat stands for Hilbert transform [30]. Using Eqn.1, along with speech representations based on sum of sine waves [18, 19], a modulation vector is now defined as

$$\tilde{\boldsymbol{M}}_k = (a_k, f_k, b_k, a_{ck}, f_{ck}, b_{ck}, p_k)^T , \qquad (2)$$

where a_k, f_k, and b_k, denote amplitude, frequency, and bandwidth parameters, which model $s_k[n]$'s spectral envelope; b_{ck} is the bandwidth around f_{ck}; and p_k is $s_k[n]$'s pitch. The relationship between f_k and f_{ck} may be understood from [21]; parameters modeling $\alpha_k[n]$ and $\beta_k[n]$ may be added using modulation spectrum [31, 32] and sub-space [33] related concepts.

Next, the elements of $\tilde{\boldsymbol{M}}_k$ are transformed, so that their scales and regions of interest, match the ones used in auditory systems [4, 12, 34], as follows: for $i = k$ and ck, a_i is converted to decibel (dB) using $10\log_{10} a_i$; f_i and b_i are mapped to the Mel scale using $2595\log_{10}(1 + F_{\text{Hz}}/700)$ [4]; a_i and b_i are capped at 200 dB and 400 Mel respectively; and p_ks outside the pitch range of 80-300 Hz are excluded. These features finally form the modulation vector, $\boldsymbol{M}_k$.

3. Travellingwave Filter Bank (TFB)

The TFB algorithm estimates and tracks $\boldsymbol{M}_k$s, by drawing inspiration from the travellingwave on the basilar membrane in the human ear's cochlea [34, 35]. Its ability to separate individual resonances, along with its hybrid representation, makes TFB superior to the spectrogram, for speech analysis.

Each channel of TFB (Fig.1) consists of a Dynamic Tracking Filter (DTF), whose feed-back loop includes a *first-order* Linear Prediction (LP) estimator [30] and a Non-linear Masker (NM). The DTF is preceded by an All Zero Filter (AZF), and coupled to a Modulation Feature Estimator (MFE). A non-linear encoder (NE) finally outputs $\boldsymbol{M}_k$ as per section 2. The basic

10.21437/Interspeech.2023-68

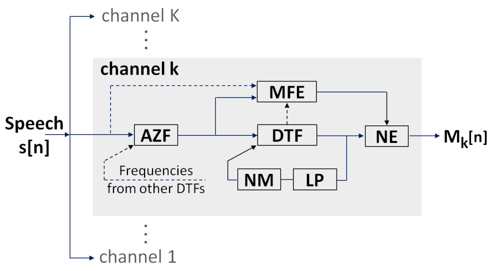

Figure 1: *TFB's resonance localized modulation tracking*

idea behind TFB is that each channel's AZF-DTF combination tracks the localized resonance's frequency, and the MFE estimates (and implicitly tracks) the modulations characterizing its associated sub-band.

3.1. Dynamic Tracking Filter

The DTF proposed is an advancement to the one in [22]. It is an adaptive single-resonance filter with a transfer function

$$H_{Dk}(n,z) = \frac{1-r_p}{1-r_p e^{j2\pi f_k[n]} z^{-1}} \ , \tag{3}$$

where k is the channel number; n is the sample number; and r_p is the pole-radius. $f_k[n]$ is estimated by LP (using its pole-angle) based on the past L samples of DTF's output. The improvements made are described next.

3.1.1. *Estimation of* $a_k[n]$*,* $b_k[n]$*, and Constant-Q Option*

$a_k[n]$ is set to be $\sqrt{\sigma_{lp}^2}$, where σ_{lp}^2 is the LP error-variance, and $b_k[n] \approx -(\ln r_{lp}) f_s/\pi$ [22], where r_{lp} is the LP pole-radius; f_s is the sampling frequency. Further, L can be made smaller, as k increases, to maintain a constant-Q [4] window. This will enable rapid and finer analysis at higher frequencies.

3.1.2. *Implementation for real-valued signals*

The DTF is implemented using the difference function,

$$s_k[n] = c_k[n] s_k[n-1] + r_p^2 s_k[n-2] + g_k[n] \tilde{s}_k[n] \ , \tag{4}$$

where $\tilde{s}_k[n]$ is the input to the DTF, $s_k[n]$ is the DTF's output, and $c_k[n]$=$2r_p$cos($2\pi f_k[n]$); the DTF's gain at $f_k[n]$ is set to unity by $g_k[n]$=$(1-r_p)\sqrt{1+r_p^2-2r_p\cos(4\pi f_k[n])}$. It avoids computation of the analytic signal [15], thereby overcoming Hilbert transform related problems [36].

3.1.3. *Non-linear Masker*

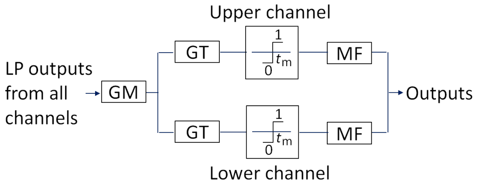

Figure 2: *NM's masking when frequencies get close*

The LP outputs from all channels are analyzed by NM (Fig.2) as follows: Get Masker (GM) sorts $f_k[n]$s and gets the strongest unmasked channel, k_s. Then Get Thresholds (GTs) compute δF_L=$f_{k_s}[n]-f_{k_s-1}[n]$ and δF_U=$f_{k_s+1}[n]-f_{k_s}[n]$ for the lower and upper channels respectively. By comparing δF_L and δF_U to a masking threshold, t_m, masking indicators, M_{iL} and M_{iU}, are computed next; set to be 0 (if $\delta F_{L/U}< t_m$) or 1 (if $\delta F_{L/U}\geq t_m$). The Masking Filters (MFs) finally output

$$f_{k_s-1}[n] = M_{iL} f_{k_s-1}[n] + (1-M_{iL}) f_{k_s-1}[n-1] \ ,$$

$$f_{k_s+1}[n] = M_{iU} f_{k_s+1}[n] + (1-M_{iU}) f_{k_s+1}[n-1]. \tag{5}$$

This process is repeated until there are no unmasked channels.

NM eliminates errors due to switching of frequency tracks. Also, it weights the frequency estimates at n-1 and n, using the estimated masking thresholds. This ensures stability of the overall (TFB) filter bank, when the DTF frequencies come close to each other. It is different from the one in [23] that sets a limit to the maximum allowable frequency spacing between DTFs, which results in tracking errors.

3.2. All Zero Filter

The transfer function for the k-th channel AZF is [22]

$$H_{Ak}(n,z) = G_k[n] \prod_{\substack{l=1 \\ l\neq k}}^{K-1} (1 - r_z e^{j2\pi f_l[n]} z^{-1}) \ , \tag{6}$$

where r_z is the radius of the AZF's zero, $f_l[n]$ is the frequency of its zero-location (obtained from other DTFs), and

$$G_k[n] = \frac{1}{\displaystyle\prod_{\substack{l=1 \\ l\neq k}}^{K-1} (1 - r_z e^{j2\pi(f_l[n]) - f_k[n]})} \tag{7}$$

normalizes the k-th DTF's gain. The improvements made to AZF include *stability (due to NM)* and *ability to handle real-valued signals*. The latter results from AZF's design using a cascade of $K-1$ filters with the l-th cascade implemented as

$$\tilde{s}_{kl}[n] = \frac{s_{kl}[n] - \tilde{c}_{kl}[n] s_{kl}[n-1] + r_z^2 s_{kl}[n-2]}{g_{kl}[n]} \ , \tag{8}$$

where $s_{kl}[n]$ is the l-th cascade's input ($s_{k1}[n]$=$s[n]$), $\tilde{s}_{kl}[n]$ is the output ($\tilde{s}_k[n]$ being the same as $\tilde{s}_{kl}[n]$ for l = K-1), $\tilde{c}_{kl}[n] \ = \ 2r_z\cos(2\pi f_l[n])$, and the normalizing gain factor is $g_{kl}[n] \ = \ \sqrt{1+r_z^2-2r_z\cos(2\pi(f_l[n]+f_k[n]))} \ \times \sqrt{1+r_z^2-2r_z\cos(2\pi(f_l[n]-f_k[n]))}$, with $g_{kl}[n] > 0$.

3.3. Modulation Feature Estimator

The k-th MFE derives a *non-distorted* sub-band spectrum, $S_{kn}[f]$, by utilizing the spectrum, $S_n[f]$, of the past L_p samples of $s[n]$ (computed only once $\forall k$, using the Fourier Transform [4]), along with left and right frequency band-edges,

$$f_{kL}[n] = \arg\min_f S_n^E[f] \left\{ \begin{matrix} f_{k-1}[n]<f<f_k[n](\forall k\neq 1) \\ 0<f<f_1[n] \end{matrix} \right\} \text{ and}$$

$$f_{kR}[n] = \arg\min_f S_n^E[f] \left\{ \begin{matrix} f_k[n]<f<f_{k+1}[n](\forall k\neq K) \\ f_K[n]<f<f_s/2 \end{matrix} \right\} \tag{9}$$

respectively; where $S_n^E[f]$ is $S_n[f]$'s spectral envelope [4, 30]. Since $f_k[n]$ is being tracked, this results in an implicit tracking of $S_{kn}[f]$. $b_{ck}[n]$ is then set to be $f_{kR}[n] - f_{kL}[n]$. $a_{ck}[n]$ and

$f_{ck}[n]$ are subsequently estimated as $a_{ck}[n] = \max |S_{kn}[f]|$ and $f_{ck}[n] = \arg \max_f |S_{kn}[f]|$.

Pitch, $p_k[n]$, is computed using $\tilde{s}_{kM}[n]$ (past L_p samples), $S_{kn}[f]$, and a hybrid of known techniques [4]. Using $p_k[n]$s and a full-band pitch estimate, $p_f[n]$, a sub-band pitch indicator, $P_k[n]$, is then defined as $P_k[n]$=1 if $p_k[n] = p_f[n]$; and 0 otherwise. As will be seen in section 4, the $P_k[n]$s yield useful cues; not provided by existing methods that group non-resonance sub-band pitches to yield one global pitch [37].

4. Modulation Synchrony

Based on several observations of $\boldsymbol{M}_k[n]$, using mixed language, gender, and age speakers, it is clear that: the simultaneous evolution of $\boldsymbol{M}_k[n]$'s elements (i.e. their synchrony), within and across channels, trace symbols that map phonemes. This "modulation synchrony" is now demonstrated using the fricative consonant, SH, having the vowel IY as its context.

For ease of explanation, and since traces of f_k and f_{ck} are similar for IY-SH-IY, let us restrict $\boldsymbol{M}_k$ to $(a_k, f_k, b_k, p_k)^T$. Also, instead of using $M_{k1}[n]$, $M_{k2}[n]$, etc., let us use a_k, f_k, b_k, and p_k; that way $a_4[t2]$-$a_1[t2]$ may be easily interpreted as amplitude difference between channels 4 and 1 at time $t2$.

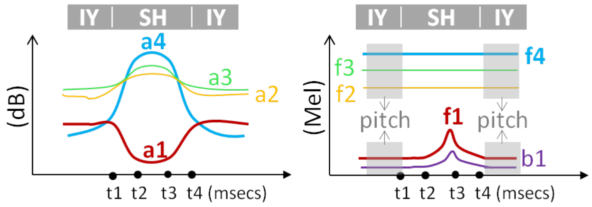

Figure 3: *Ideal symbol for SH (with IY on left and right)*

Fig.3 displays the **acoustic symbol** that has been observed for IY-SH-IY. Let R^r=$t2$:$t3$ denote SH's resonance region. The details of cues in Fig.3 are then as follows.

For resonance: a_4 exceeds a_1 by at least 19 dB; the maximum range of all a_ks is at least 3 dB greater than the maximum range of a_2, a_3, and a_4; SH's peak amplitude is greater than those of its adjoining IYs; f_2 and f_1 are above 1500 and 250 Mel respectively; only b_1 is above 125 Mel (b_2, b_3, b_4 are below 125 Mel); all P_ks are absent for SH; and R^r's duration is between 30 and 500 msecs. And for transition: durations ($t1$:$t2$ and $t3$:$t4$) are between 10 and 100 msecs, and a_4's rise and drops are greater than 5 dB. These (acoustic) cues may be expressed as

$$a_4[v] - a_1[v] \geq t_a^1 \; (v \in R^r) \tag{10}$$

$$w_{1:4}^a[v] \div w_{2:4}^a[v] \geq t_a^2 \; (v \in R^r \; , \; w_{2:4}^a[v] \neq 0) \tag{11}$$

$$a_{max} - a_{max}^- \geq t_a^3 \; , \; a_{max} - a_{max}^+ \geq t_a^3 \tag{12}$$

$$f_2[v] \geq t_f^1 \; , \; f_1[v] \geq t_f^2 \; (v \in R^r) \tag{13}$$

$$b_1[v] \geq t_b^1 \; , b_j[v] \leq t_b^1 \; (j = 2, 3, 4 \; ; \; v \in R^r) \tag{14}$$

$$P_j[v] = 0 \; (\forall j \; , \; v \in R^r) \tag{15}$$

$$t_d^1 \leq (t3 - t2) \leq t_d^2 \tag{16}$$

$$t_d^3 \leq (t2 - t1) \leq t_d^4 \; , \; t_d^3 \leq (t4 - t3) \leq t_d^4 \tag{17}$$

$$a_4[t2] - a_4[t1] \geq t_s^1 \; , \; a_4[t3] - a_4[t4] \geq t_s^1 \; ; \tag{18}$$

where $w_{i:j}^a[v]$ represents the synchrony of a_i to a_j at v, using max $(a_i[v], a_{i+1}[v], ..., a_j[v])$-min $(a_i[v], a_{i+1}[v], ..., a_j[v])$; a_{max}, and a_{max}^-, and a_{max}^+, are the maximum values of a_ks, for SH, left IY, and right IY respectively; the thresholds t_a^j, t_f^j, t_b^j, t_d^j, and t_s^j (j=1,2,...) can be estimated using standard statistical [30] or deep learning [38] techniques. Earlier studies [14] that characterize SH by dominant high frequency energy, relative amplitude, and noise duration, have reported only cues that are similar to Eqns. 13, 12 , and 16 respectively.

The set of cues in Eqns. 10:18 form the **acoustic code** for IY-SH-IY. Eqns. 10 and 11 that correspond to predominant features of the symbol in Fig. 3, which are necessary to characterize the phoneme, are called the **main cues**; and $a_4[v]$-$a_1[v]$, $w_{1:4}^a[v] \div w_{2:4}^a[v]$ are called **main cue-features**.

5. Simulations

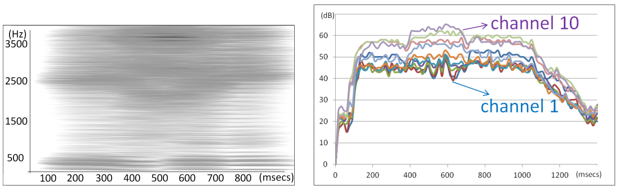

Figure 4: *Spectrogram (Left) and MFB Outputs (Right)*

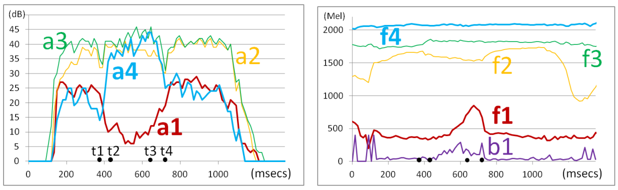

Figure 5: a_k*s (Left) and* f_k*s*+b_1 *(Right)*

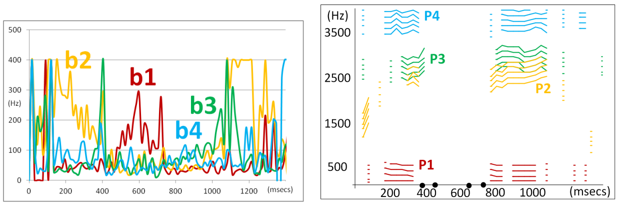

Figure 6: *All* b_k*s (Left) and Sub-band Harmonic Tracks (Right)*

First, results of analyzing an utterance corresponding to IY-SH-IY, spoken by a male speaker, using a Motorola Z2 Force smart-phone, are presented. TFB parameters were r_p=0.9, r_z=0.99, L=120, t_m=250, and Lp=256; K=4 and f_s=8 KHz.

For this example, the spectrogram (widely used for acoustic-phonetic mapping [14]) is shown in Fig.4 Left, and outputs of the Mel Filter Bank (MFB), which is the *de facto* standard for speech recognition feature extraction [4, 1], is shown in Fig. 4 Right. Apart from high frequency energy, found in many phonemes, they fail to yield other cues, specific to SH.

Other problems associated with them include: a) peak-picking the spectrogram or choosing the right MFB channels, to track resonances, is not trivial [24, 25, 26, 27], b) any chosen MFB filter's center frequency, may not line up with the signal's resonance, resulting in frequency estimation errors, and c) MFB's triangular weighted averaging could bias estimates of cues based on energies - e.g., energy difference between the two "manually selected" channels (1 and 10), whose center frequencies are close to 1st and 4th formant locations, is only ≈ 20 dB, as opposed to the true value of ≈ 36 dB (computed manually).

In contrast, Fig.5 displays an entire set of cues that form an acoustic symbol, similar to Fig.3. Specifically, Fig.5 (Left) shows that at peak resonance (t_R=540 msecs), a_4-a_1=36 dB; and a_2, a_3, a_4 are grouped together relative to their separation from a_1 ($w^a_{1:4}$÷$w^a_{2:4}$=18 dB). Further, $a_{max}>\left(a^-_{max}, a^+_{max}\right)$, and a_4's transitions during rise (375:435 msecs) and drop (645:720 msecs) are steep (23 dB and 19 dB respectively). Fig.5 (Right) shows that in the resonance region, $f_2>$ 1500 Mel, $f_1>$ 250 Mel, and b_1>100 Mel. Also, the values of f_2, f_3, and f_4 are similar to those of their adjoining IYs.

Fig.6 (Left) displays all b_ks for this example. Notice that only b_1 exhibits deviations during SH's resonance. Further, observe that Fig.6 (Right) shows no harmonic lines (no p_ks) for any of SH's resonances, whereas IYs display mostly all P_ks.

Clearly, the example considered maps to the acoustic code of Eqns. 10:18, with thresholds: t^1_a=36, t^2_a=18, t^3_a=3, t^1_f=1500, t^2_f=250, t^1_b=100, t^1_d=t^2_d=400, t^3_d=60, t^4_d=70, and t^1_s=19.

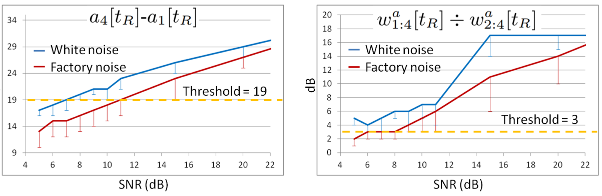

Figure 7: *Robustness of main cue-features in noise*

In Fig. 7, the average (μ) of main cue-features, sampled at t_R, as a function of signal-to-noise ratio (SNR = $10\log_{10}\left(\frac{P_s}{\sigma^2}\right)$, where P_s is the speech power with *silence excluded*, and σ^2 is the noise power) is plotted; error-bars indicate $-\sigma$. A comparison of $\mu-\sigma$ to thresholds, reveals that TFB is robust at SNRs ≥8 dB for white noise; at lower SNRs, at least one cue-feature's $\mu-\sigma$ falls below threshold, and the symbol looses its predominant shape. For factory noise, due to its intermittent bursts, σ is relatively higher and TFB is robust only for SNRs ≥15 dB. Thus, TFB has potential to extract symbols even in noise.

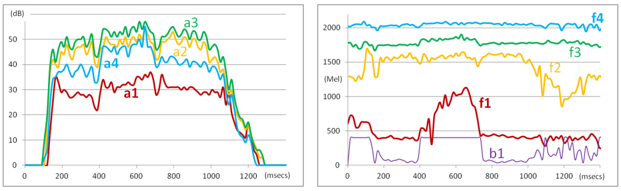

Figure 8: a_k*s (Left) and* f_k*s+*b_1 *(Right) for* $r_p = 0.1$

The effect of increasing TFB's DTF bandwidth is shown in Fig. 8. A comparison of Fig. 8 and Fig. 5, shows that TFB is not very sensitive to the choice of r_p. However, at very low values, a_ks and f_ks display relatively more fluctuations (due to energy leakage from other sub-bands), and the symbol gets distorted. On decreasing DTF bandwidths, TFB will fail to track f_ks and not yield symbols; similar to fixed filter banks like MFB.

Finally, Table 1 lists the number of additions, multiplications, and total calculations, needed for TFB (with 4 channels) and MFB (with 10 channels); only the filtering algorithms for TFB and MFB are compared; for TFB, only DTF, NM, and AZF (shown in Fig. 1) are considered; and for MFB, the computation of FFT, Mel cepstrum, delta cepstrum, and delta-delta cepstrum [4], are ignored. As can be seen, MFB requires ≈ 25 times more computations every second, compared to TFB. Results of further analysis, more examples, and links to TFB source-code and data-sets (to enable reproducibility), are in [39].

Table 1: *Comparison of* 4 *channel TFB with* 10 *channel MFB*

# Computations / sec	TFB	MFB
Additions	1139	29949
Multiplications	1273	29949
Total	2412	59898

6. Discussions and Future Work

The acoustic symbols and codes derived for all English language phonemes (documented in [39]) indicate that a) the latter may be mapped to *unique* context-dependent shapes (similar to Fig. 3) and machine-readable rules (similar to Eqns. 10:18), which the spectrogram and MFB fail to accomplish; and b) all acoustic cues reported in earlier studies [14] correlate well, but with only a sub-set of cues rendered by the symbols. However, experiments reveal that some of the code equations (e.g. Eqns. 12:14 and 18, for IY-SH-IY) are not always satisfied for speakers enunciating poorly [39]; reinforcing the challenge of variability in speech. Interestingly, the code structure resembles layers of linear transforms coupled with non-linearities, seen in deep learning neural networks [40, 38]. For instance, Eqn. 10 is a linear combination of $a_4[v]$ and $a_1[v]$, followed by a non-linearity ($> t^1_a$), where each a_k is output of linear filters (convolutional AZF and recurrent DTF in Fig. 1) that is non-linearly transformed (by NE in Fig. 1). These new insights may be used to estimate code thresholds, in a way that the resulting codes enable speech recognition systems to require lesser training data, and be more robust to training-testing model mismatch [2, 3].

Further, using the code equations, a confidence metric may be defined as $\mathbf{C}_{ac}$=$\frac{\text{\#Matching-Cues}}{\text{\#Total-Cues}}\times 100$. It may be extended to word levels, and subsequently used for enabling speech understanding [41] and multi-modal [42] systems, to generate feedback, such as: display choices when 77%≤$\mathbf{C}_{ac}$<100%, prompt "speak clearly" if 55%≤$\mathbf{C}_{ac}$<77%, and prompt "please repeat" for $\mathbf{C}_{ac}$<55%. Even further, the codes may be used to perform advanced voice analytics [43]. For example, t^1_a=30 in Eqn. 10 indicates that SH was spoken loudly; t^2_a=5 in Eqn. 11 indicates that SH was enunciated clearly; and t^3_a=0 in Eqn. 12 implies that IY was louder than its following SH phoneme.

TFB's design around just 4 channels, each using simple (1-pole DTF and 3-zeros AZF) filters, makes it highly attractive for low-cost hardware and software implementations. The fine-tuning of its 5 parameters may be viewed as time-frequency filtering [15] "matched" to the acoustic symbols of phonemes.

The time-alignments that form part of the acoustic symbols (e.g., $t1$, $t2$, $t3$, $t4$, in Fig. 3), are currently being manually computed. An algorithm to automatically estimate these is work in progress. It will also facilitate more detailed acoustic-phonetic analysis, across multiple languages, using a large data-set.

7. Conclusion

The synchrony in speech modulation vectors, estimated using TFB's resonance localized tracking, helps derive acoustic symbols and codes that map phonemes. The codes have potential to improve many aspects of current speech recognition systems.

8. References

[1] X. Huang, J. Baker, and R. Reddy, "A historical perspective of speech recognition," *Communications of the ACM*, vol. 57, no. 1, pp. 94–103, Jan. 2014.

[2] B. S. Atal, "Automatic speech recognition: A communication perspective," *Proceedings of the IEEE ICASSP*, vol. 1, pp. 457–460, May 1999.

[3] A. Rao, B. Roth, V. Nagesha, D. McAllaster, N. Liberman, and L. Gillick, "Large vocabulary continuous speech recognition of read speech over cellular and landline networks," *Proceedings of the ICSLP*, pp. 402–405, Oct. 2000.

[4] L. R. Rabiner and R. W. Schafer, "An introduction to digital speech processing," *Foundations and Trends in Signal Processing*, vol. 1, no. 1-2, pp. 1–194, 2007.

[5] H. Fletcher, "The relative difficulty of interpreting the spoken sounds of English," *Physical Review*, vol. 15, pp. 413–516, Nov. 1920.

[6] G. Fant, "Half a century in phonetics and speech research," *Fonetik 2000, Swedish phonetics meeting in Skövde*, pp. 2852–2861, May 2000.

[7] N. Mesgarani, S. David, and S. Shamma, "Representation of phonemes in primary auditory cortex: How the brain analyzes speech," *Proceedings of the IEEE ICASSP*, vol. 4, pp. 765–768, May 2007.

[8] A. Lahiri and H. Reetz, "Distinctive features: Phonological underspecification in representation and processing," *Journal of Phonetics*, vol. 38, pp. 44–59, Jan. 2010.

[9] J. B. Allen, "How do humans process and recognize speech?," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, Oct. 1994.

[10] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychological Review*, vol. 74, pp. 431–461, May 1967.

[11] S. E. Iblumstein and K. N. Stevens, "Phonetic features and acoustic invariance in speech," *Cognition*, vol. 10, no. 1, pp. 25–32, 1981.

[12] J. B. Allen and F. Li, "Speech perception and cochlear signal processing," *IEEE Signal Processing Magazine*, vol. 26, pp. 73–77, July 2009.

[13] F. Li, A. Trevino, A. Menon, and J. B. Allen, "A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise," *J. of the Acous. Soc. of America*, vol. 132, pp. 2663–2675, Oct. 2012.

[14] H. Reetz and A. Jongman, *Phonetics: Transcription, Production, Acoustics, and Perception*, John Wiley and Sons, Hoboken, New Jersey, 2020.

[15] L. Cohen, *Time Frequency Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, 1995.

[16] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. of the Acous. Soc. of America*, vol. 50, pp. 637–655, Aug. 1971.

[17] A. Katsiamis, E. Drakakis, and R. Lyon, "Practical gammatone-like filters for auditory processing," *EURASIP Journal on Audio, Speech, and Music Processing*, Dec. 2007, 063685 (2007).

[18] J. L. Flanagan, "Parametric coding of speech spectra," *J. of the Acous. Soc. of America*, vol. 68, pp. 412–419, Aug. 1980.

[19] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3024–3051, Oct. 1993.

[20] H. B. Voelcker, "Toward a unified theory of modulation - Part I: Phase envelope relationships," *Proceedings of the IEEE*, vol. 54, no. 3, pp. 340–354, Mar. 1966.

[21] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *J. of the Acous. Soc. of America*, vol. 105, no. 3, pp. 1912–1924, Mar. 1999.

[22] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 240–254, May 2000.

[23] K. Mustafa and I. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Transactions on Speech and Audio Processing*, pp. 435–444, Apr. 2006.

[24] J. L. Flanagan, "Automatic extraction of formant frequencies from continuous speech," *J. of the Acous. Soc. of America*, vol. 27, pp. 110–118, Jan. 1955.

[25] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. of the Acous. Soc. of America*, vol. 47, no. 2, pp. 634–648, Feb. 1970.

[26] B. S. Atal and C. H. Shadle, "Decomposing speech into formants: A new look at an old problem," *J. of the Acous. Soc. of America*, vol. 64, pp. S162, Nov. 1978.

[27] L. B. Jackson and J. Bertrand, "An adaptive inverse digital filter for formant analysis of speech," *Proceedings of the IEEE ICASSP*, pp. 84–86, Apr. 1976.

[28] A. Rao and R. Kumaresan, "Dynamic tracking filters for decomposing nonstationary sinusoidal signals," *Proceedings of the IEEE ICASSP*, pp. 917–920, May 1995.

[29] V. K. Peddinti, R. Kumaresan, and P. Cariani, "Improved auditory-inspired signal processing algorithm design for tracking multiple frequency components," *SN Computer Science*, vol. 1, Jan. 2020, 62 (2020).

[30] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[31] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

[32] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Advances in Signal Processing*, pp. 668–675, June 2003.

[33] D. W. Tufts and R. Kumaresan, "Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 975–989, Sept. 1982.

[34] J. O. Pickles, *An Introduction to the Physiology of Hearing, 2nd Ed.*, Academic, London, U.K., 1988.

[35] E. de Boer, "Auditory physics. Physical principles in hearing theory. III," *Physics Reports*, vol. 203, no. 3, pp. 125–231, May 1991.

[36] A. Rao and R. Kumaresan, "A parametric modeling approach to Hilbert transformation," *IEEE Signal Processing Letters*, vol. 5, pp. 15–17, Jan. 1998.

[37] B. Lee and D. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," *Proceedings of the Interspeech*, Sept. 2012.

[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016, `http://www.deeplearningbook.org`.

[39] A. Rao, "Travellingwave Filter Bank analysis and phoneme mapping using acoustic symbols and codes for the English language," *Supplementary Material for Interspeech Paper - https://travellingwave.com/Tain3231.pdf*, pp. 1–45, May 2023.

[40] O. Abdel-Hamid, A. Rahman Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[41] T. J. Hazen, S. Seneff, and J. Polifroni, "Recognition confidence scoring and its use in speech understanding systems," *Computer Speech & Language*, pp. 49–67, Jan. 2002.

[42] A. Rao, "System and method for multimodal utterance detection," *USA Patent 9922640*, 2018.

[43] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human–computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Feb. 2001.